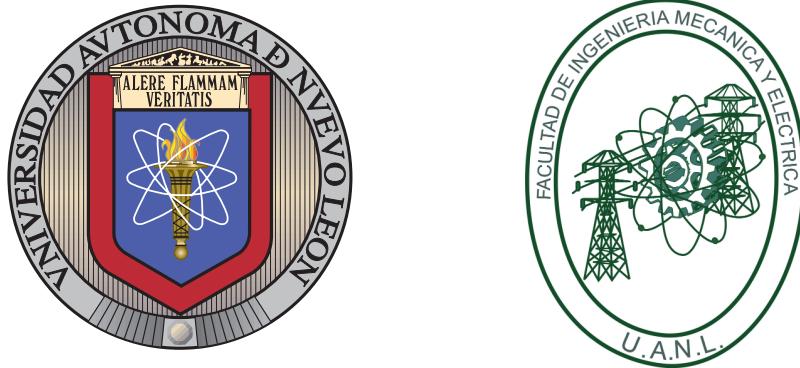


UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA
POSGRADO EN INGENIERÍA DE SISTEMAS
DOCTORADO



PORTAFOLIO DE EVIDENCIAS

DE

JOHANNA BOLAÑOS ZUÑIGA

1883900

PARA EL CURSO DE MODELOS PROBABILISTAS APLICADOS,

CON LA PROFESORA DRA. ELISA SCHAEFFER .

SEMESTRE AGOSTO 2020 - ENERO 2021.

[GITHUB.COM/JOHANNABZ/PROBABILIDAD/TREE/MASTER/PORTAFOLIO](https://github.com/johannabz/probabilidad/tree/master/portafolio)

Tarea 1

1. Recopilación de datos - Sección Precios

En el presente trabajo se realiza un análisis estadístico con el programa R versión 4.0.2. Se utilizó la información del Índice Nacional de Precios al Consumidor (INPC)¹ mensual por ciudades en el periodo de enero 2019 - 2020. Esta información fue consultada en la página del INEGI [2], en la sección de Precios.

La información de las 55 ciudades del INPC y su respectivo índice con base en la segunda quincena de julio 2018, fueron descargadas en un archivo .xlsx. Para efectos del estudio, se filtró la información de las 3 principales ciudades de México (Monterrey, Ciudad de México y Guadalajara), la cual se muestra en la Tabla 1. Estos datos se guardaron en un archivo .dat.

Tabla 1: Índice Nacional de Precios al Consumidor

Fecha	Monterrey	Ciudad De México	Guadalajara
Ene 2019	103.548	102.653	102.078
Feb 2019	103.791	102.513	102.468
Mar 2019	104.090	102.859	103.275
Abr 2019	102.993	103.291	103.646
May 2019	102.945	103.523	103.827
Jun 2019	102.998	103.741	103.893
Jul 2019	103.762	104.079	103.993
Ago 2019	103.814	103.817	104.003
Sep 2019	104.038	104.134	104.124
Oct 2019	105.636	104.372	104.543
Nov 2019	105.892	104.952	105.060
Dic 2019	106.301	105.659	105.764
Ene 2020	106.793	106.060	105.805

¹El INPC es un indicador económico que muestra la variación de los precios en un periodo de tiempo.

2. Diagrama de cajas y bigotes

Para realizar el análisis estadístico y como ayuda visual para identificar el comportamiento de los datos de la sección 1, se utilizó el diagrama de cajas y bigotes. El código en R se encuentra en el repositorio de GitHub [1]

De acuerdo a la Figura 1 podemos observar que, entre las 3 ciudades principales de México, el INPC más bajo lo reportó Guadalajara y el más alto lo registró Monterrey, mientras que, Ciudad de México registró una variación atípica. También se puede observar que la mediana de los INPC de las tres ciudades es muy similar, aunque Monterrey registró una mayor variabilidad, mantuvo por más tiempo INPC bajos, sin embargo, Guadalajara mantuvo por más tiempo los INPC altos.

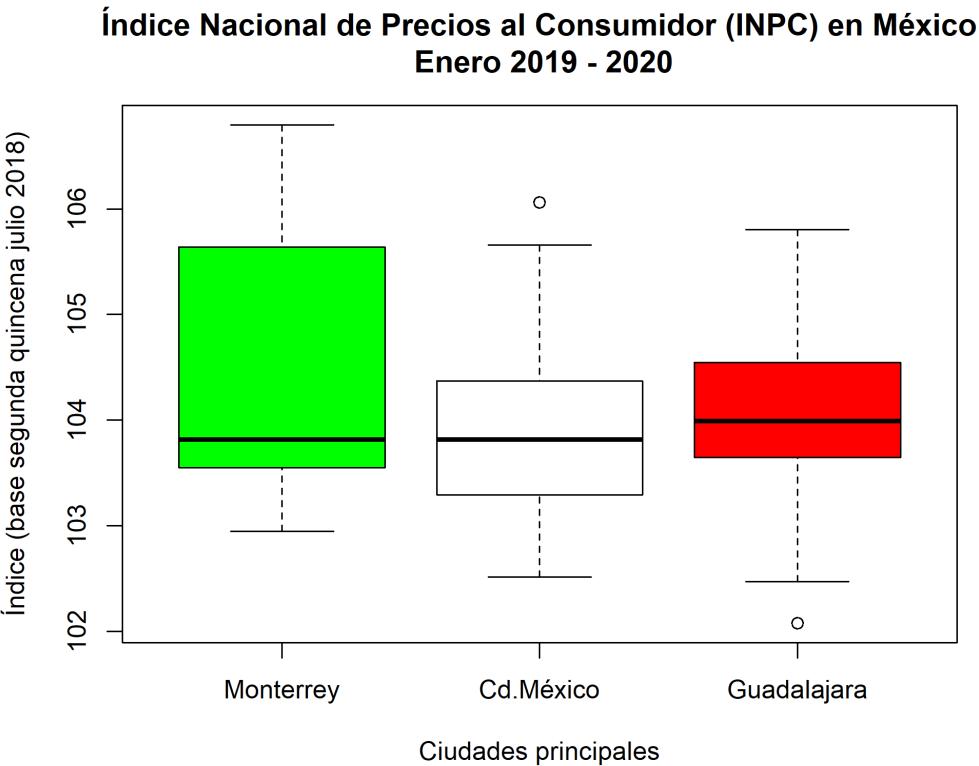


Figura 1: Índice Nacional de Precios al Consumidor de enero 2019 - 2020 en las 3 ciudades principales de México

De igual forma, en la Figura 2 se muestran los índices de las 55 ciudades del INPC reportadas por la INEGI, en la cual podemos observar que el INPC más alto lo reportó Hermosillo y el más bajo fue en Tijuana. También se puede observar que varias ciudades registraron variaciones atípicas en el INPC.

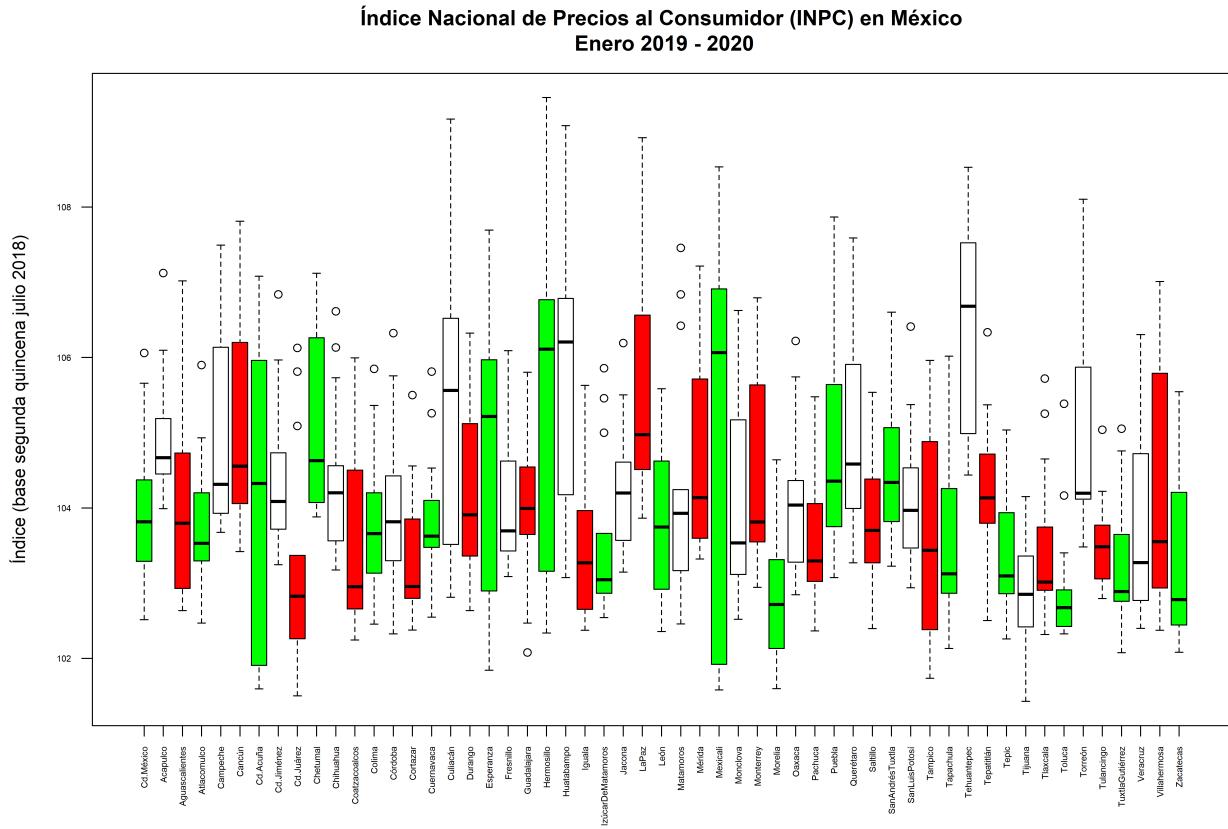


Figura 2: Índice Nacional de Precios al Consumidor de enero 2019 - 2020 en las 55 ciudades del INPC en México

Referencias

- [1] Johanna Bolaños Z. Repositorio en github de la clase de modelos probabilistas aplicados. Recurso libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea %201, 2020.
- [2] INEGI. Índice nacional de precios al consumidor (inpc). Recurso libre, disponible en <https://www.inegi.org.mx/datos/>, 2020.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 2

1. Libro Seleccionado — *The Scarlet Letter*

En el presente trabajo se realizó un análisis estadístico en el programa R versión 4.0.2 [3] utilizando el libro *The Scarlet Letter*, el cual se encuentra disponible de manera gratuita en *Project Gutenberg* [2].

2. Análisis estadístico

Para realizar el análisis estadístico y como ayuda visual se utilizaron diferentes tipos de gráficos. El código en R se encuentra en el repositorio de GitHub [1]. Para llevar a cabo el estudio, el libro fue analizado tanto por la cantidad de **letras** como de **palabras** que hay en él.

En la Figura 1, se muestran todos los caracteres alfanuméricos utilizados en el libro. Sin embargo, nos interesan todos las letras, por lo cual, se procedió a realizar un filtro y, finalmente, obtenemos los datos mostrados en la Figura 2, en la cual se puede observar todas las letras y la cantidad de veces en que fueron empleadas. Para visualizar de una mejor manera la información anterior, se organizan de manera decreciente la frecuencia de las letras, ver Figura 3.

En cuanto al análisis por *palabras*, en la Figura 4 se muestran todas las palabras usadas en el libro. Sin embargo, se puede observar que son tantas las palabras utilizadas que es difícil hacer una interpretación a partir de ella. Por lo anterior, se realizó una depuración de las palabras que aportan poca información, como son las preposiciones, artículos y conjunciones. Estas palabras son conocidas como *stopwords*.

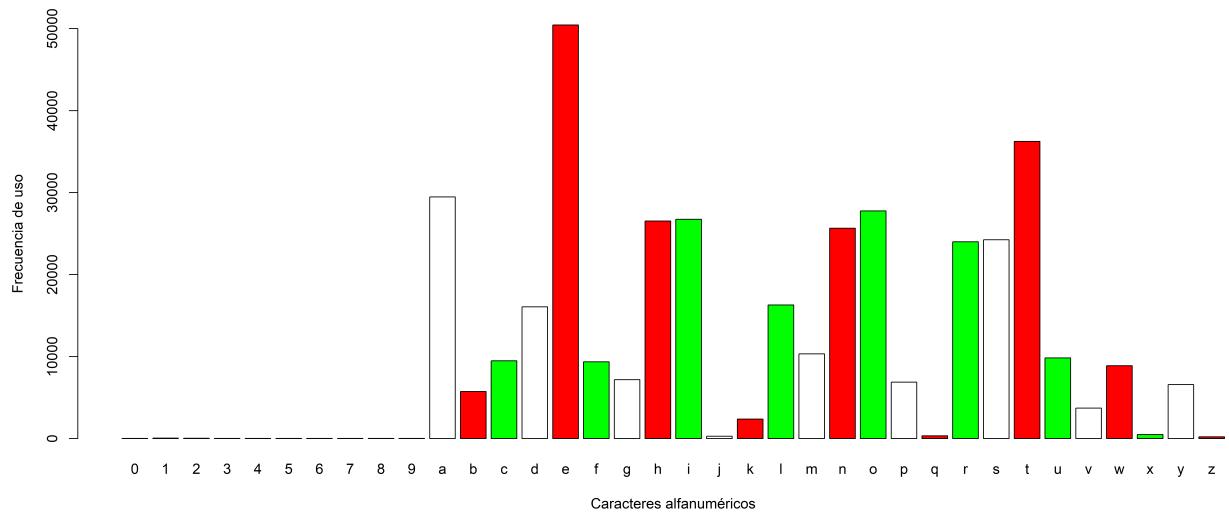


Figura 1: Frecuencia de los caracteres alfanuméricos del libro

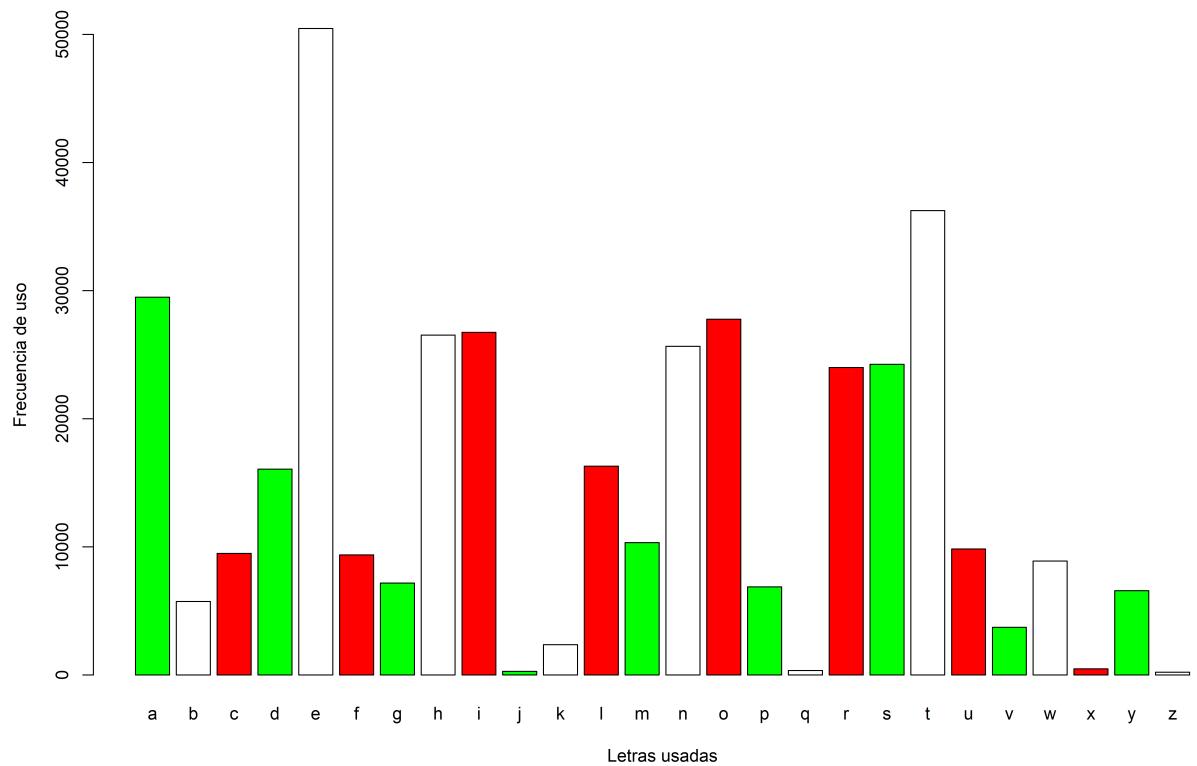


Figura 2: Frecuencia de las letras usadas en el libro.

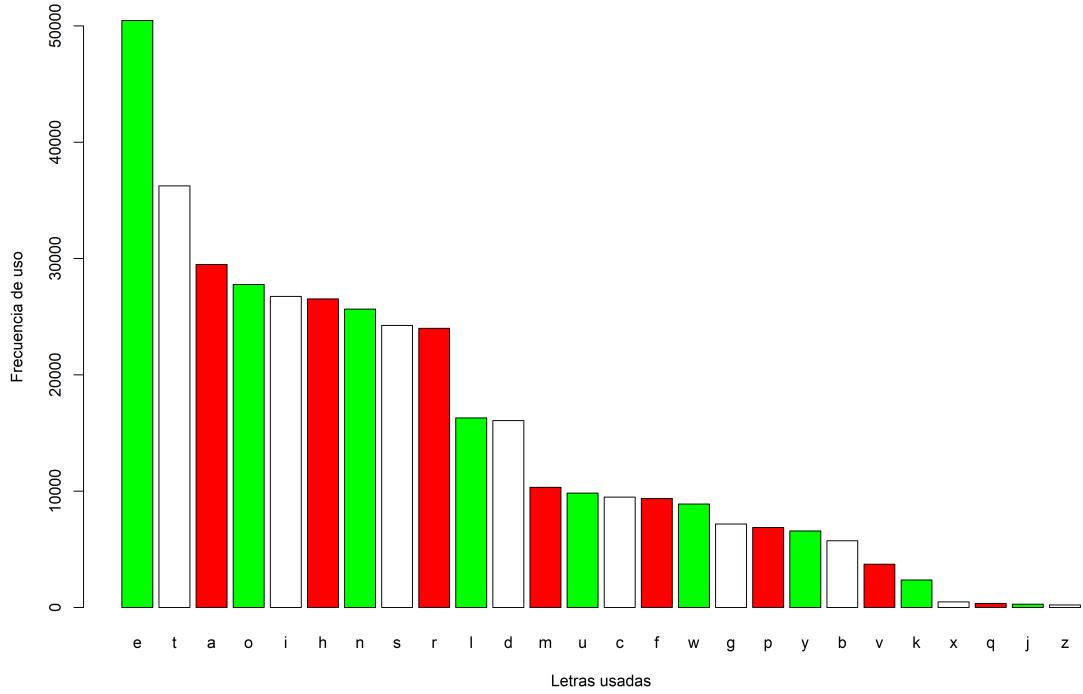


Figura 3: Frecuencia decreciente de las letras usadas en el libro.

En la Figura 5, se muestran las palabras que tienen una frecuencia superior a 150. Para una mejor interpretación de los datos obtenidos, en la Figura 6 se organizan de manera decreciente las palabras más usadas. De las palabras con mayor frecuencia podemos destacar que, la palabra más repetida es **Hester** y **Pearl**, lo cual tiene mucho sentido, ya que la primera es el nombre del personaje principal y la segunda, es el nombre de la hija. Sin embargo, a pesar de que **Prynne** es el apellido de la protagonista, el autor frecuenta más su nombre.

En la Figura 7, se muestra de una manera más didáctica las palabras más usadas en el libro, donde el tamaño es mayor para las palabras que aparecen con más frecuencia, este tipo de gráfica se conoce como **nube de palabras** o en inglés *world cloud*.

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea2, 2020.
- [2] Hawthorne, Nathaniel. The Scarlet Letter. Recurso libre, disponible en <http://www.gutenberg.org/ebooks/25344>, 1985.

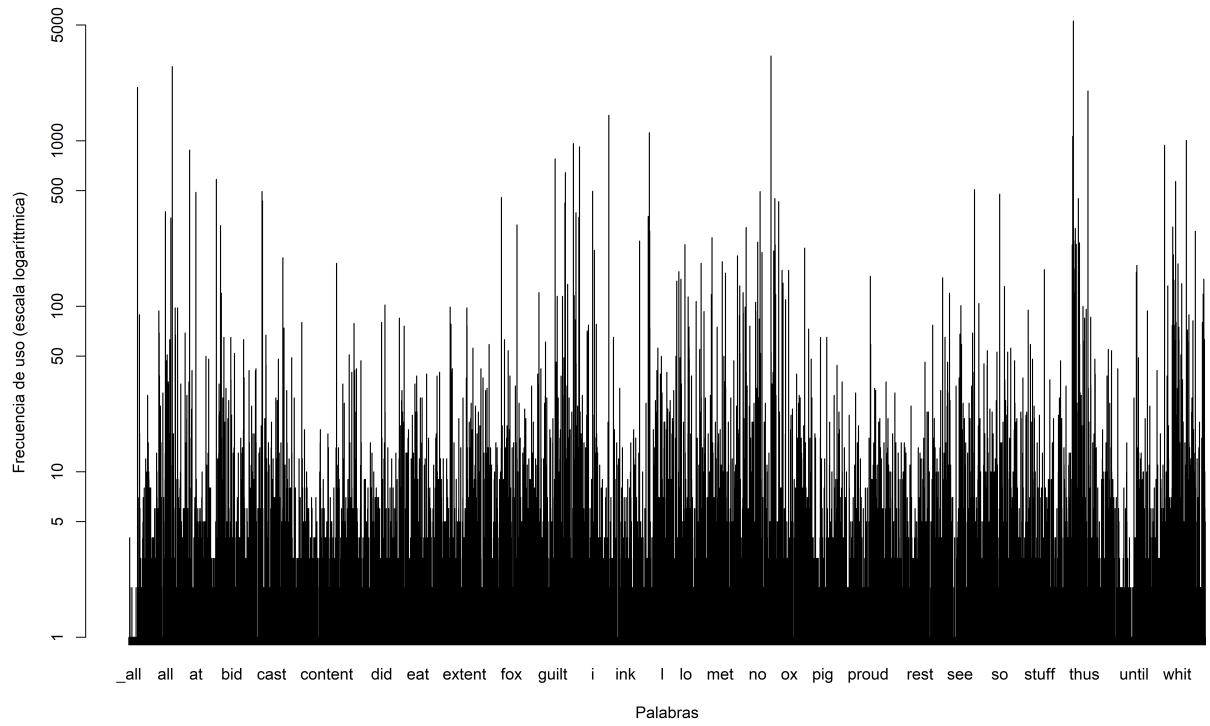


Figura 4: Palabras utilizadas en el libro

[3] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.

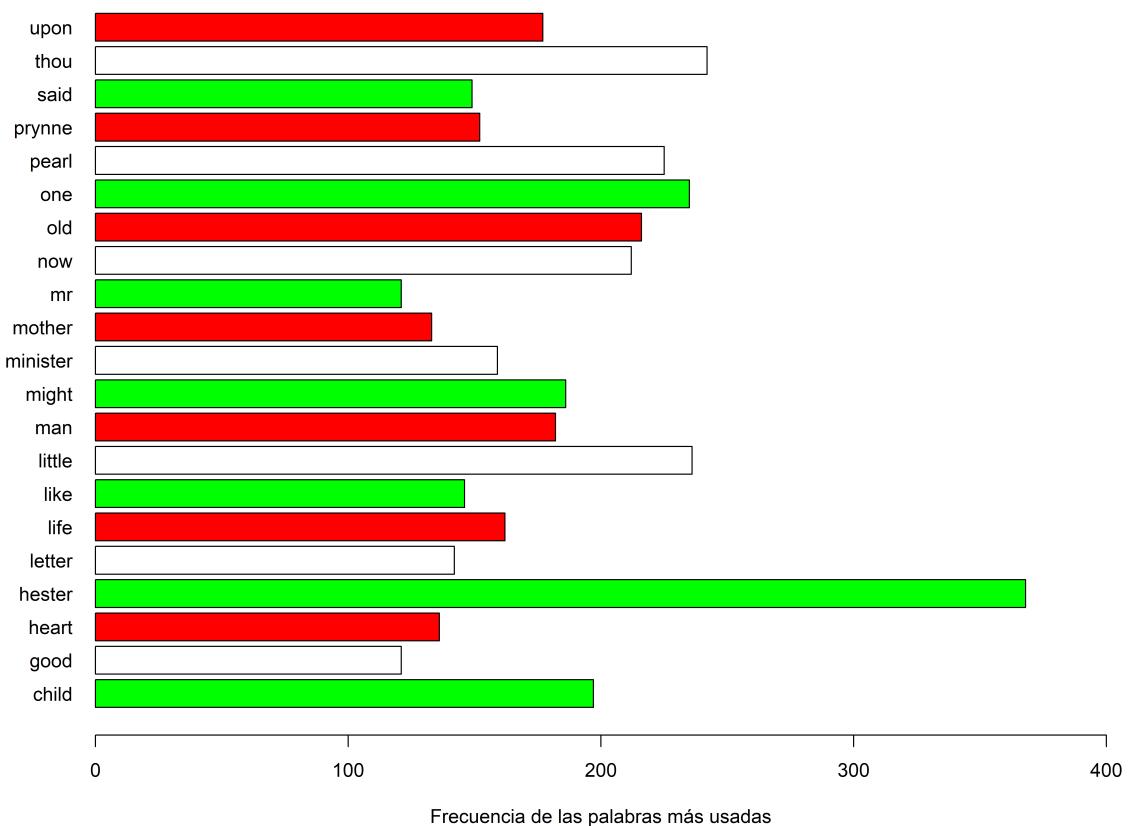


Figura 5: Palabras de mayor frecuencia en el libro

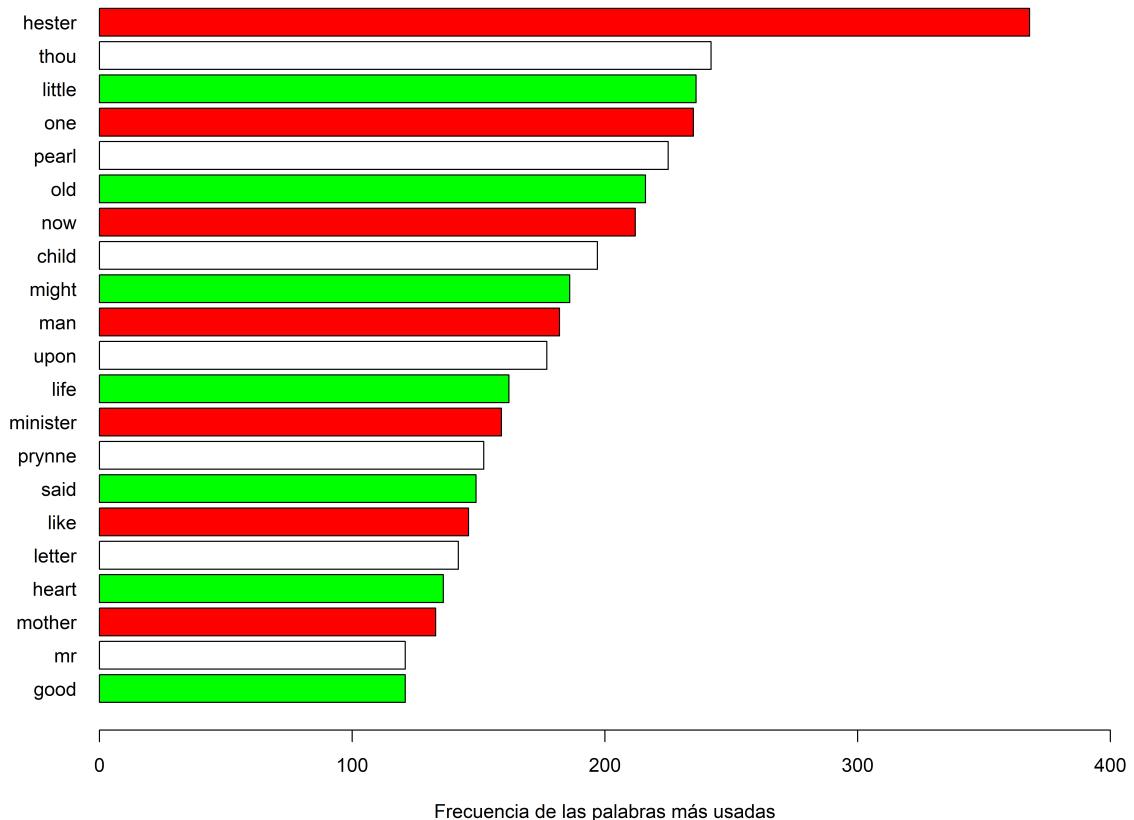


Figura 6: Frecuencia decreciente de las palabras más usadas en el libro.



Figura 7: Nube de palabras más frecuentes en el libro

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 3

1. Libro Seleccionado — *The Scarlet Letter*

En el presente trabajo se continua con el libro *The Scarlet Letter*, el cual se encuentra disponible de manera gratuita en *Project Gutenberg* [2]. En esta ocasión, se analizó el tipo de distribuciones discretas que podrían estar contenidas en el escrito mediante el programa R versión 4.0.2 [3]. El código en R se encuentra en el repositorio de GitHub [1].

2. Distribuciones

Para el análisis de las distribuciones, se consideró reflejar cómo se comporta la longitud de las palabras presentes en el libro y la cantidad de palabras en cada párrafo. Para este último criterio, se determinó usar párrafos que contienen más de 9 palabras, ya que el título más largo contiene 9 palabras.

De acuerdo a la figura 1, se puede observar que la distribución de la longitud de cada palabra se asemeja a una distribución binomial. Mientras que, en la figura 2 correspondiente a la cantidad de palabras por párrafos, es similar a la geométrica.

3. Simulación de distribuciones

Con el fin de simular las distribuciones geométrica, binomial y binomial negativa, se escogieron como criterios algunas características relevantes del análisis estadístico realizado anteriormente, como por

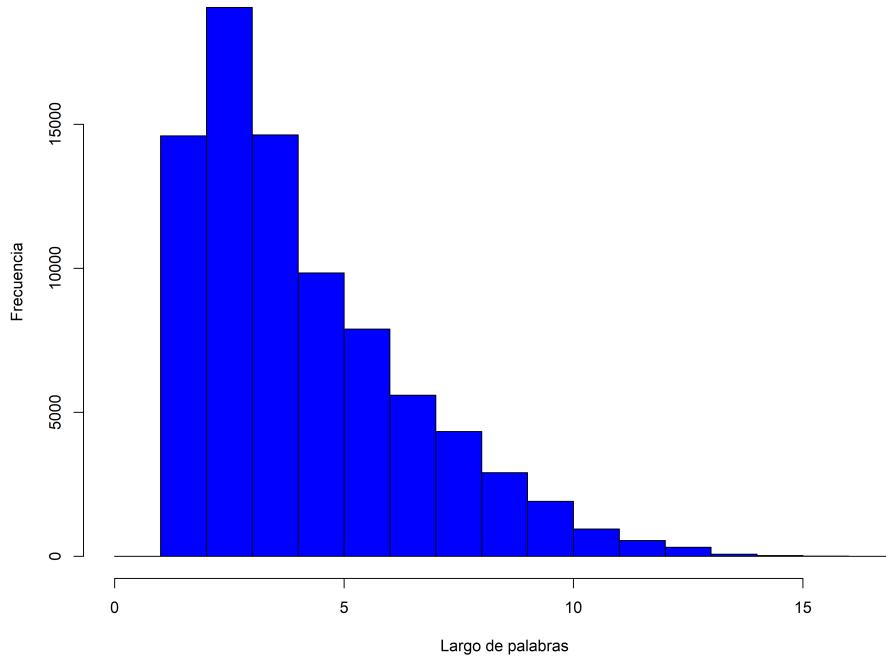


Figura 1: Distribución de la longitud de las palabras usadas en el libro

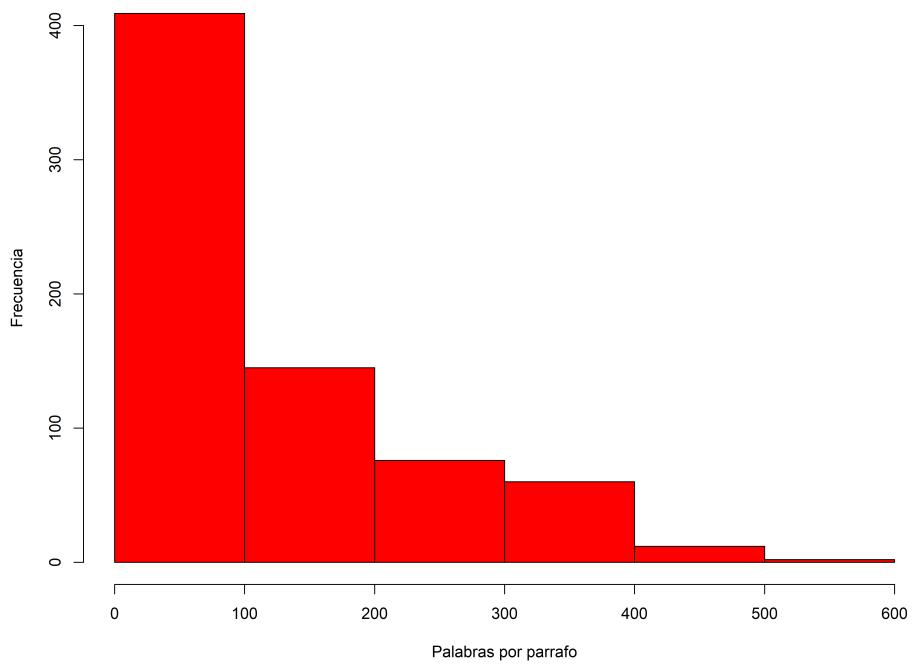


Figura 2: Distribución de la cantidad de palabras en cada párrafo (en párrafos con más de 9 palabras)

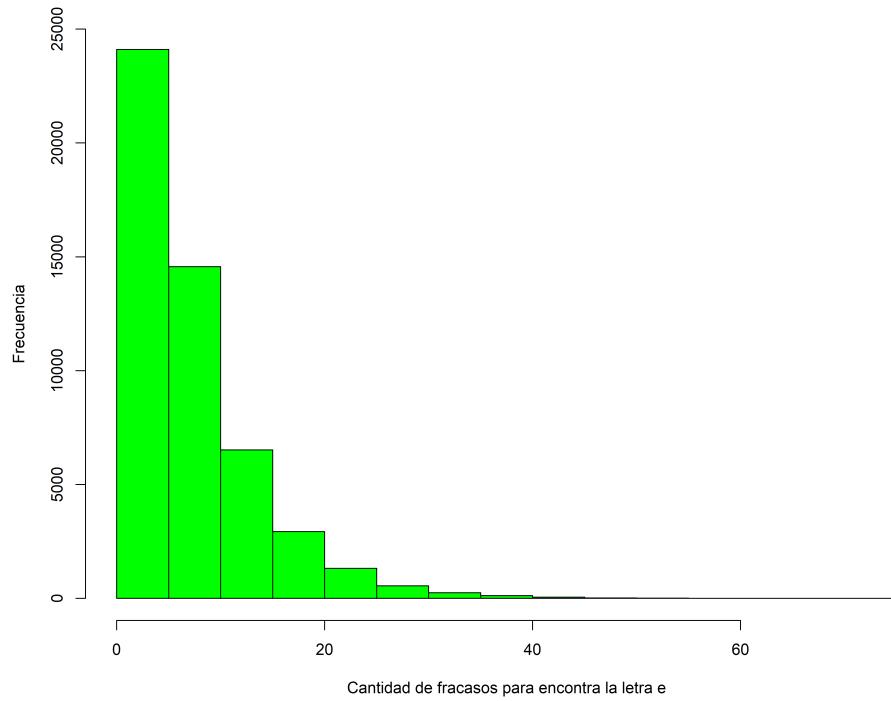


Figura 3: Distribución geométrica usando como éxito la letra con mayor frecuencia en el libro

ejemplo la letra y palabra con mayor frecuencia.

Para la distribución geométrica, se consideró el criterio de cuantos caracteres alfanuméricos hay antes de encontrar la letra con mayor frecuencia en todo el texto, en este caso, el éxito será encontrar la letra **e**. En la figura ref se puede observar como el criterio seleccionado sigue una distribución geométrica.

Para la distribución binomial se consideró que las repeticiones se formarían con los párrafos cuya longitud de palabras es igual a 9, ya que son los más frecuentes y como longitud de palabra igual a 2, ya que es la menor cantidad de caracteres que pueda contener este tipo de párrafos. Con esta simulación se pretende encontrar cuantas palabras con la longitud dada (éxitos) se encuentran en los párrafos con las características mencionadas anteriormente. En la figura 4 se puede observar como el criterio seleccionado sigue una distribución binomial. Cabe mencionar que como la muestra es muy pequeña, son muy pocos los datos obtenidos.

Finalmente, para la distribución binomial negativa, se consideró hallar la cantidad de intentos para completar 5 éxitos en todo el libro, se considera un éxito cuando la letra **h** tiene seguida la letra **e**. Se contempló este criterio ya que la palabra y artículo más frecuente es **hester** y **the**, respectivamente. En la figura 5 se puede observar como el criterio seleccionado sigue una distribución binomial negativa.

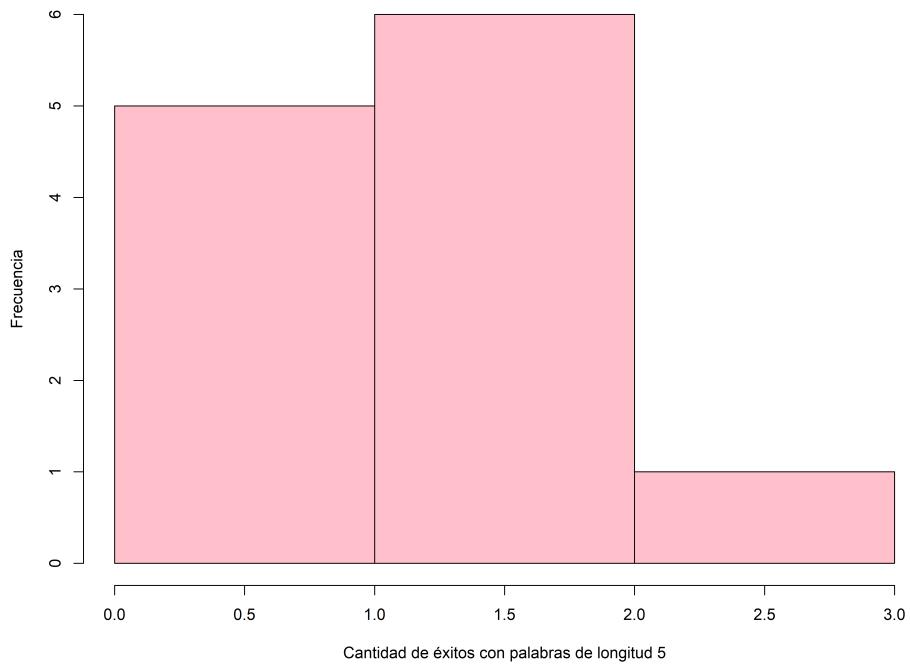


Figura 4: Distribución binomial usando como éxito la cantidad de palabras con una determinada longitud

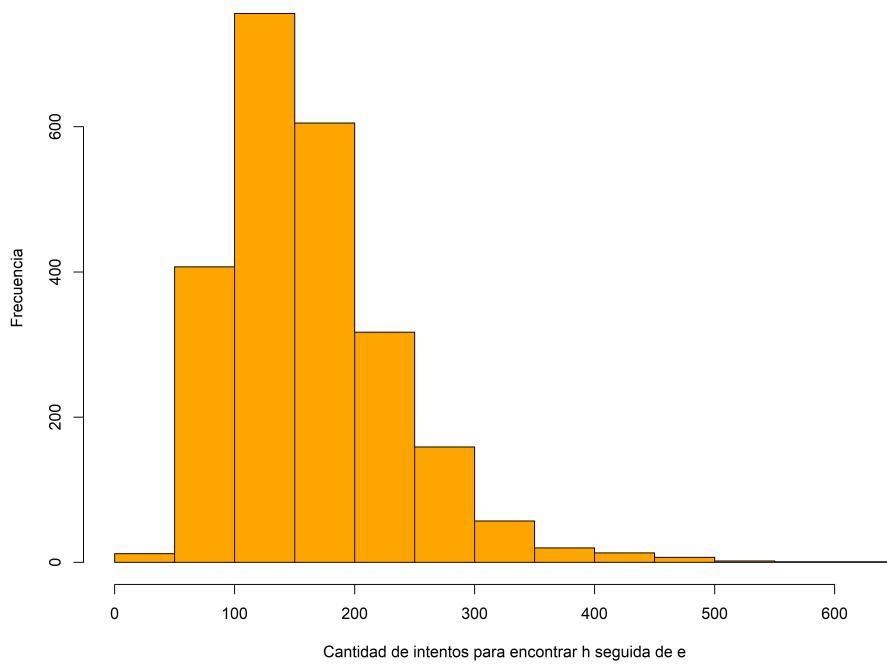


Figura 5: Distribución binomial negativa

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea3, 2020.
- [2] Hawthorne, Nathaniel. The Scarlet Letter. Recurso libre, disponible en <http://www.gutenberg.org/ebooks/25344>.
- [3] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 4

1. Introducción

En el presente trabajo se presenta un análisis de la relación que existe entre la distribución Poisson y las distribuciones exponencial y uniforme. Este análisis es complementado con resultados obtenidos de experimentos computacionales realizados en el software R versión 4.0.2 [7]. El código empleado se encuentran en el repositorio GitHub [1].

2. Distribución de Poisson

El proceso de Poisson, por lo general, se utiliza en escenarios donde se cuentan las ocurrencias de ciertos eventos que parecen ocurrir a un determinado ritmo, pero completamente al azar. Con base en Walpole [8], la distribución de Poisson se utiliza para calcular la probabilidad obtener “eventos” o llegadas durante un periodo particular. Es una distribución de probabilidad discreta donde la variable aleatoria (v.a) de Poisson X representa el número de eventos que ocurren en un intervalo unitario. La ecuación 1, representa la probabilidad de la v.a Poisson, donde lambda λ es la tasa media constante conocida (por unidad de tiempo) e independiente del tiempo transcurrido desde el último evento.

$$p(x; \lambda) = \frac{e^{-\lambda}(\lambda)^x}{x!}, \quad x = 0, 1, 2, \dots, \quad (1)$$

3. Distribución exponencial y uniforme

De acuerdo a Walpole [8], la distribución exponencial desempeña un papel importante en la teoría de colas y en problemas de confiabilidad. Los tiempos entre llegadas en instalaciones de servicio y los tiempos de operación antes de que las partes de los componentes y sistemas eléctricos empiecen a fallar, suelen representarse mediante la distribución exponencial. El significado de la variable aleatoria se puede interpretar como el tiempo que transcurre hasta que se produce un fallo, la probabilidad de que el elemento, la pieza o el componente considerado dure cierta cantidad de tiempo. Esta distribución es un caso particular de distribución gamma con alfa $\alpha = 1$. Su parámetro se denota la letra lambda λ . En la ecuación 2 y 3 se muestran la función de densidad de la distribución exponencial y la acumulativa, respectivamente [4].

$$f(x) = \lambda e^{-\lambda x} \quad (2)$$

$$F(x) = 1 - e^{-\lambda x} \quad (3)$$

En una distribución uniforme, todos los posibles eventos x tienen la misma probabilidad de ocurrencia. Esta distribución se define tanto para el caso discreto como el continuo [4]. La variable aleatoria solo puede tomar valores comprendidos entre dos extremos a y b , de manera que todos los intervalos de una misma longitud (dentro de (a, b)) tienen la misma probabilidad. En la ecuación 4 y 5 se muestran las funciones de densidad absoluta y acumulativa de distribución uniforme (continua), respectivamente [4, 8].

$$f_{(x)} = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x > b \end{cases} \quad (4)$$

$$F_{(X)} = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases} \quad (5)$$

4. Relación de la distribución exponencial y uniforme con el proceso de Poisson

De acuerdo al estado del arte y a los conceptos desarrollados en las secciones anteriores, la relación que se puede deducir respecto a la distribución uniforme con el proceso de Poisson es que, en un proceso de Poisson de parámetro λ (número de eventos por unidad de tiempo), los instantes en los cuales ocurren los eventos sucesivos dividen el intervalo de tiempo t en una sucesión de intervalos disjuntos. Entonces, como han ocurrido n eventos o llegadas, la probabilidad de que un evento haya ocurrido en uno de esos intervalos es la misma, por lo tanto, se puede decir que exhibe una distribución uniforme. La demostración matemática de esta relación, se puede encontrar en [6, 8, 5]

De igual manera, se pudo deducir que la relación respecto a la distribución exponencial con el proceso de Poisson es que, el tiempo que transcurre hasta que suceda de un evento o llegada (o el tiempo entre llegadas) siguen una distribución exponencial, por lo cual la cantidad de llegadas es lo mismo que la cantidad de valores exponencialmente distribuidos requeridos para completar un valor unitario [4] La demostración matemática de esta relación, se puede encontrar en [6, 8].

En la literatura [5, 3, 2], se encuentran los algoritmos que ayudan simular a partir de las distribuciones uniforme y exponencial, el comportamiento de la distribución Poisson. Con base en estos algoritmos, se realizaron experimentos numéricos para determinar los parámetros en que las distribuciones se asemejan entre sí.

4.1. Experimentación numérica

Para el caso de la distribución de Poisson a partir de una distribución exponencial, el algoritmo va sumando los valores obtenidos de una distribución exponencial y va contando cuantos valores tuvo que generar para alcanzar una *meta*, la cual fue fijada en 1. Los valores de λ y el número de replicas n que se realiza este experimento fueron variando.

De igual forma, Para el caso de la distribución de Poisson a partir de una distribución uniforme, el algoritmo va multiplicando los valores obtenidos de una distribución uniforme (intervalo $[0, 1]$) y va contando cuantos valores generó hasta que ese producto sea menor a $e^{-\lambda}$. Los valores de λ y n fueron variando.

En la figura 1, 2 y 3 se muestran los resultados de la experimentación cuando las réplicas son de 100, 1,000 y 50,000 con un $\lambda = 2$, $\lambda = 20$ y $\lambda = 70$, respectivamente. Los valores para la distribución de Poisson fueron generados con la función `rpos(n, λ)`.

De acuerdo a la forma que van tomando los histogramas de las figuras 1, 2 y 3. se llega a la

conclusión de que a medida que aumenta la cantidad de replicas n y el valor de λ se hace más grande, se obtiene una similitud entre las distribuciones.

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea4, 2020.
- [2] Devroye, Luc. *Non-uniform random variate generation*. Springer-Verlag New York Inc., 1986.
- [3] Knuth, Donald Ervin. *Seminumerical Algorithms, The Art of Computer Programming*. Addison Wesley, third edition, 1997.
- [4] Schaeffer, Elisa. Modelos probabilistas aplicados: notas del curso. Recurso disponible en, <https://elisa.dyndns-web.com/teaching/prob/pisis/prob.htmlut2>.
- [5] Sigman, Karl. Inverse Transform Method. Recurso disponible en, <http://www.columbia.edu/ks20/4404-Sigman/4404-Notes-ITM.pdf>.
- [6] Taylor, Howard M., Karlin, Samuel. *An Introduction to Stochastic Modeling*. Academic Press, third edition, 1998.
- [7] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [8] Walpole, Ronald E., Myers, Raymond H., Myers Sharon L., Ye, Keying. *Probability statistics for Engineers Scientists*. Pearson Education, Inc., 9th edition, 2012.

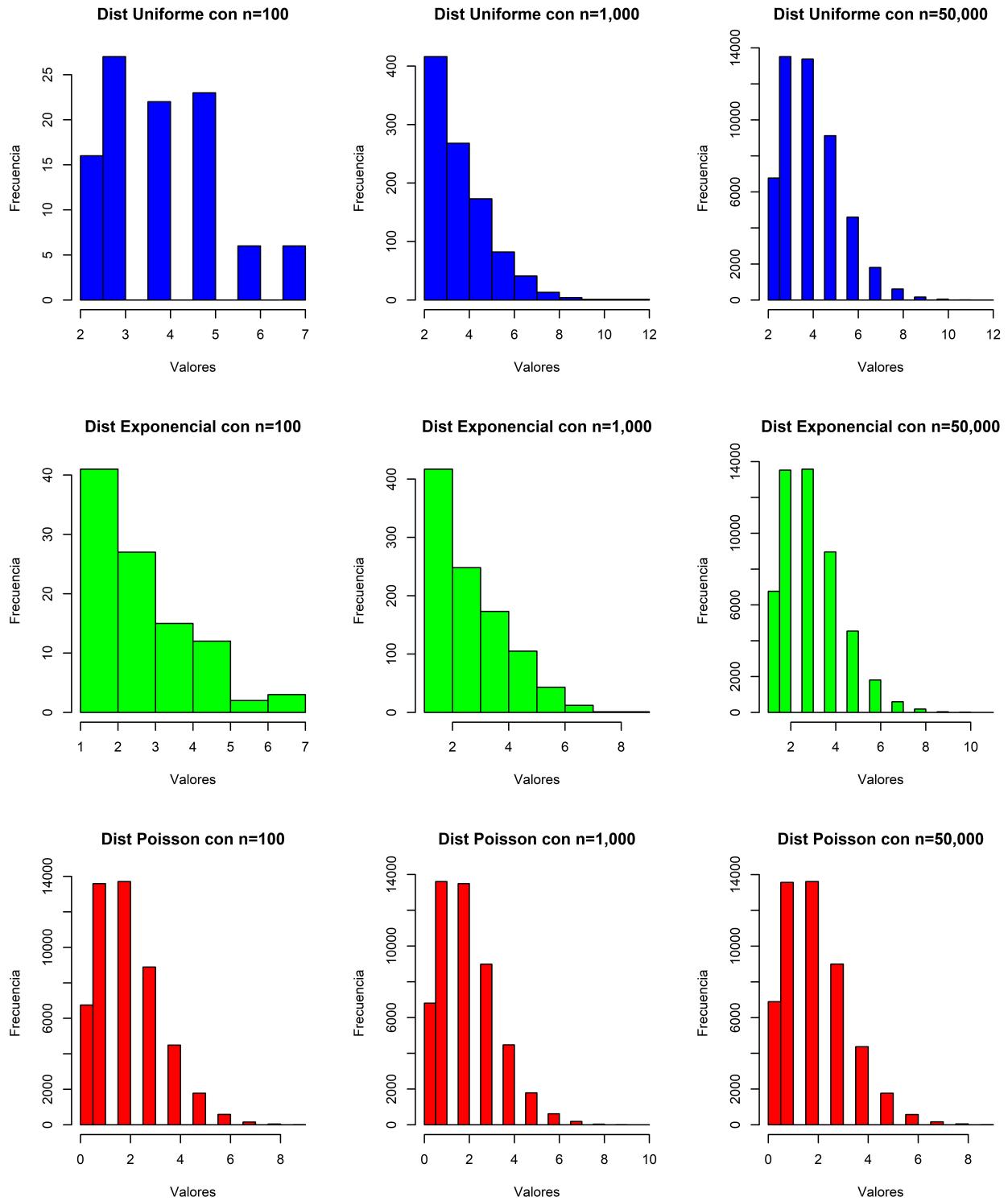


Figura 1: Distribución de Poisson (rojo) a partir de la distribución uniforme (azul) y exponencial (verde) con $\lambda = 2$

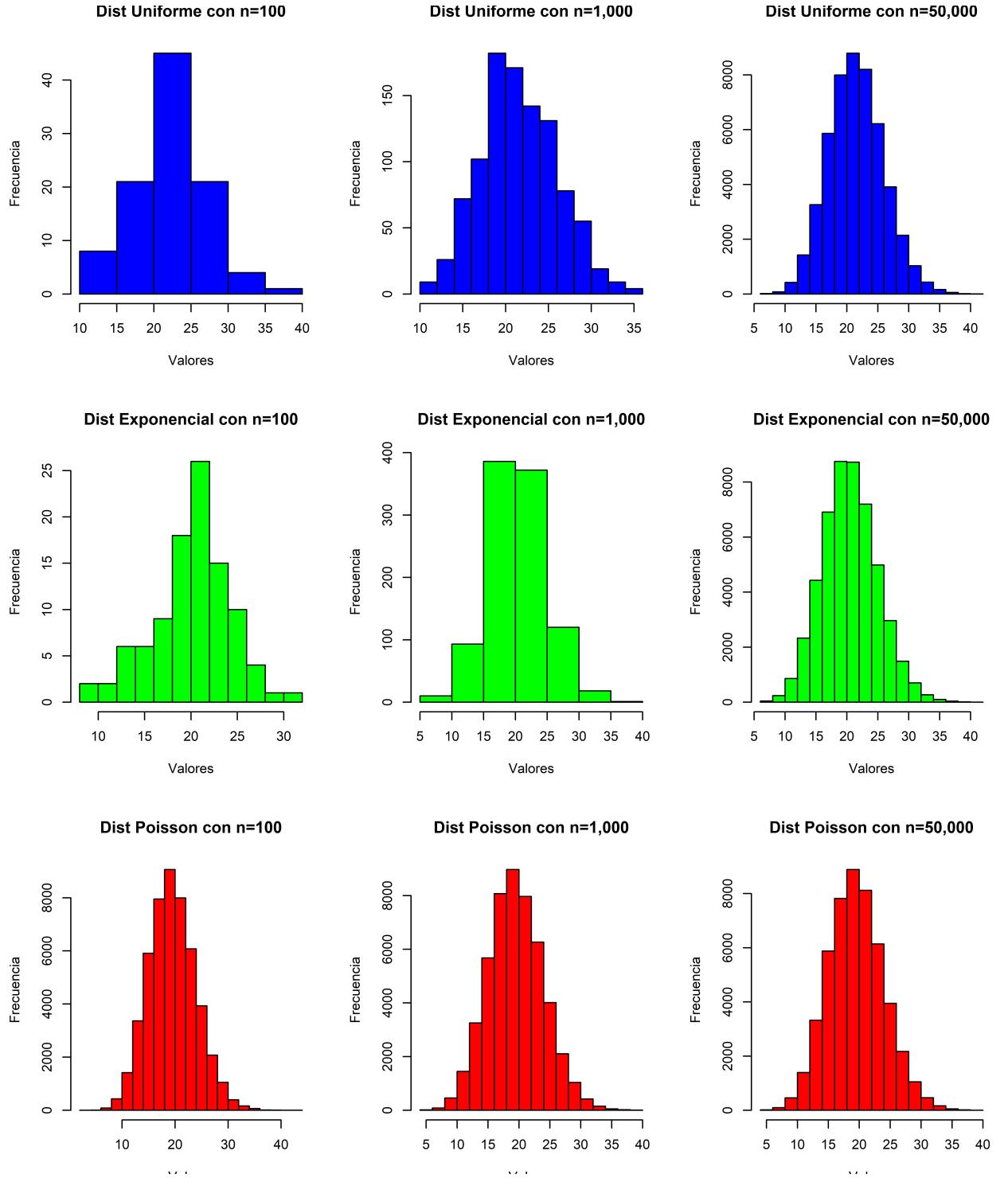


Figura 2: Distribución de Poisson (rojo) a partir de la distribución uniforme (azul) y exponencial (verde) con $\lambda = 20$

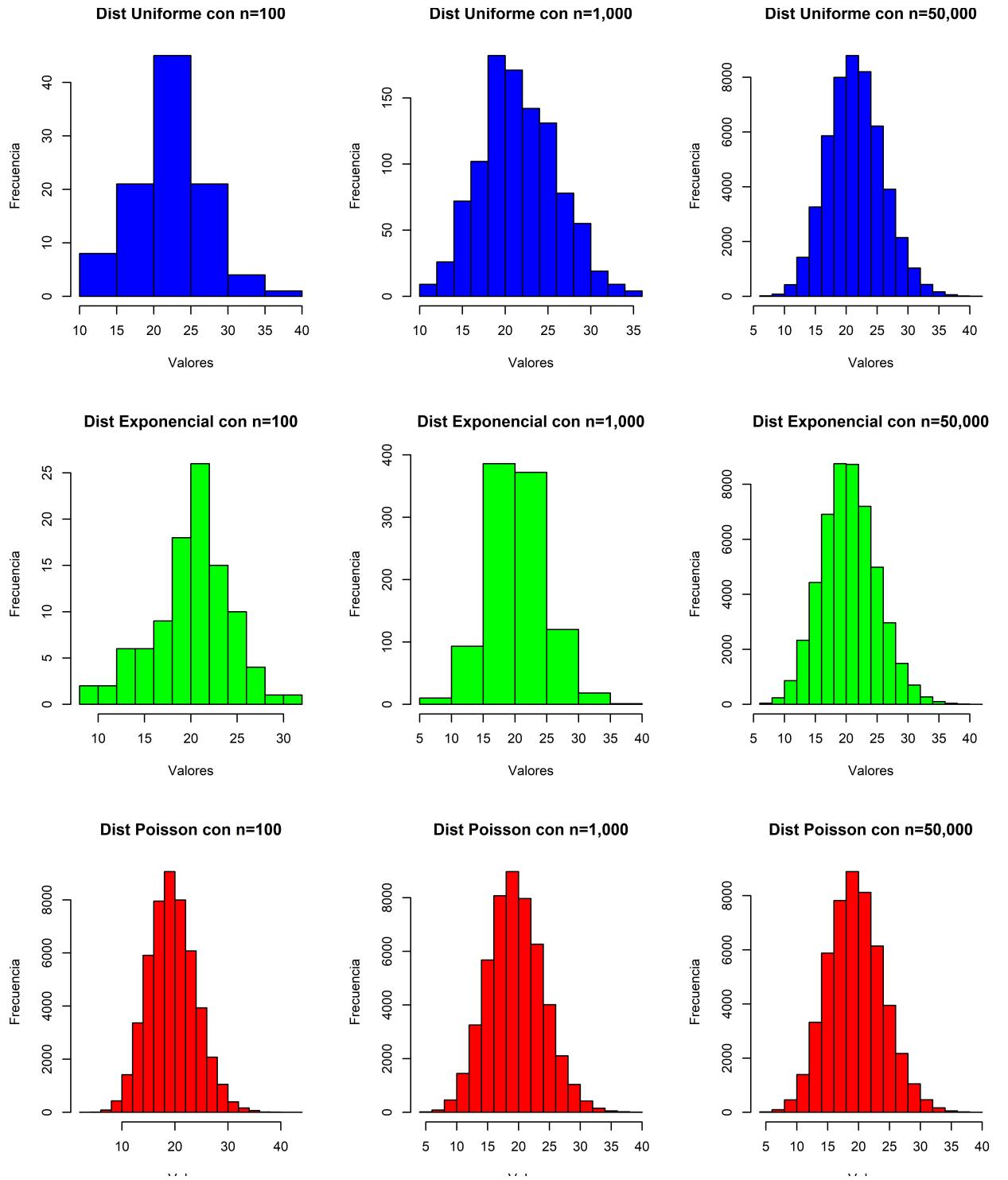


Figura 3: Distribución de Poisson (rojo) a partir de la distribución uniforme (azul) y exponencial (verde) con $\lambda = 70$

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 5

1. Introducción

En este trabajo se presentan algunos algoritmos para la generación pseudoaleatoria de números con distribución normal y uniforme. Se utilizan también pruebas de normalidad y uniformidad para verificar que los datos generados siguen este tipo de distribuciones. Se trabaja en el software R versión 4.0.2 [7] y el código empleado se encuentran en el repositorio GitHub [1].

2. Generación de números pseudoaleatorios uniformes

De acuerdo a Hiller [5] los números generados por una computadora no se deben llamar números aleatorios porque son predecibles y se pueden reproducir, dado el generador de números aleatorios que se use. Por ello, en ocasiones se les llama números pseudoaleatorios. Los números pseudoaleatorios son los generados por medio de la computadora (algoritmo), el cual es un proceso que parece producir números al azar, pero no lo hace realmente. Casi cualquier plataforma computacional sabe generar enteros pseudoaleatorios dentro de un rango predeterminado desde cero hasta un máximo (que suele ser una potencia de dos). A partir de ellos, se puede generar otro tipo de valores pseudoaleatorios con aritmética simple [6].

Para la generación de números pseudoaleatorios uniformes (entre $(0, 1)$) a los cuales denotaremos con r_i , independientemente del algoritmo o procedimiento que se utilice para su generación, lo importante son los números que genera, ya que estos deben poseer ciertas características que aseguren su validez [3]. Dichas características son:

-
- Deben estar uniformemente distribuidos.
 - Ser estadísticamente independientes.
 - Reproducibles.
 - Su periodo o ciclo de vida debe ser largo.
 - Deben ser generados a través de un método que no requiera mucha capacidad de almacenamiento de la computadora.
 - Se inicia con una semilla X_0 y, a partir de esta, se van generando X_1, X_2, X_3, \dots

Dada la importancia de contar con un conjunto suficientemente grande de números pseudoaleatorios, existen diferentes algoritmos determinísticos para obtenerlos. En este trabajo nos enfocaremos en el algoritmo generador congruencial mixto, ya que es el más utilizado. Este algoritmo genera una secuencia de números enteros por medio de la siguiente ecuación recursiva:

$$X_{i+1} = (aX_i + c) \bmod m \quad i = 0, 1, 2, \dots, N \quad (1)$$

donde a (constante aditiva), c (constante multiplicativa) y m (módulo o longitud del ciclo) son enteros positivos, X es la secuencia de valores pseudoaleatorios y X_0 es la semilla. Como se mencionó anteriormente, los números que se generan son enteros, por lo cual, de acuerdo a García [4], para obtener números pseudoaleatorios uniformes se requiere de la ecuación 2

$$r_i = \frac{X_i}{m - 1}. \quad (2)$$

A modo de ejemplo, se consideró generar $n = 8$ números pseudoaleatorios con los siguientes parámetros: $X_0 = 100$, $a = 6$, $c = 22$ y $m = 30$. La secuencia de los números pseudoaleatorios uniformes r_i se muestran en la tabla 1. Para determinar si los datos obtenidos del generador siguen una distribución uniforme, se utilizó la función `uniform.test`. Después de aplicar la prueba, se obtuvo que $p\text{-value} = 0,945$ es mayor que 0,05 lo que significa que los datos obtenidos al parecer siguen una distribución uniforme.

De la anterior experimentación se pueden realizar las siguientes inferencias: en el cuadro 1, podemos observar que el número generado en X_3 y en X_8 son los mismos y que al aplicar la prueba de uniformidad nos arrojó una advertencia, lo cual podría indicar que, de acuerdo a García [4], si continuamos generando más números con los parámetros dados, estos se repetirían. Para comprobar lo anterior, se realizó otra experimentación donde se generaron 30 números pseudoaleatorios con los mismos parámetros iniciales, y al aplicar la prueba de uniformidad con el comando `uniform.test`, el $p\text{-value} = 0,002292$ fue menor que 0,05, lo que significa que los datos obtenidos al parecer no siguen una distribución uniforme.

Cuadro 1: Números pseudoaleatorios uniformes del generador $X_{i+1} = (6X_i + 22) \bmod 30$

i	X_i	r_i
1	22	0,7586207
2	4	0,1379310
3	16	0,5517241
4	28	0,9655172
5	10	0,3448276
6	22	0,7586207
7	4	0,1379310
8	16	0,5517241

```
Chi-squared test for given probabilities

data: hist.output$counts
X-squared = 0.75, df = 4, p-value = 0.945

Warning message:
In chisq.test(x = hist.output$counts, p = probs) :
  Chi-squared approximation may be incorrect
```

[test.txt]

```
Chi-squared test for given probabilities

data: hist.output$counts
X-squared = 24, df = 8, p-value = 0.002292

Warning message:
In chisq.test(x = hist.output$counts, p = probs) :
  Chi-squared approximation may be incorrect
```

De acuerdo a Hiller [5], García [4] y Coss [3], para que el generador de números pseudoaleatorios uniforme sea eficiente, se requiere que los parámetros X_0 , a , c y m cumplan ciertas condiciones y sugieren lo siguiente:

- $m = p^d$, donde p es la base del sistema (binario, decimal, hexadecimal, etc.) que se está utilizando

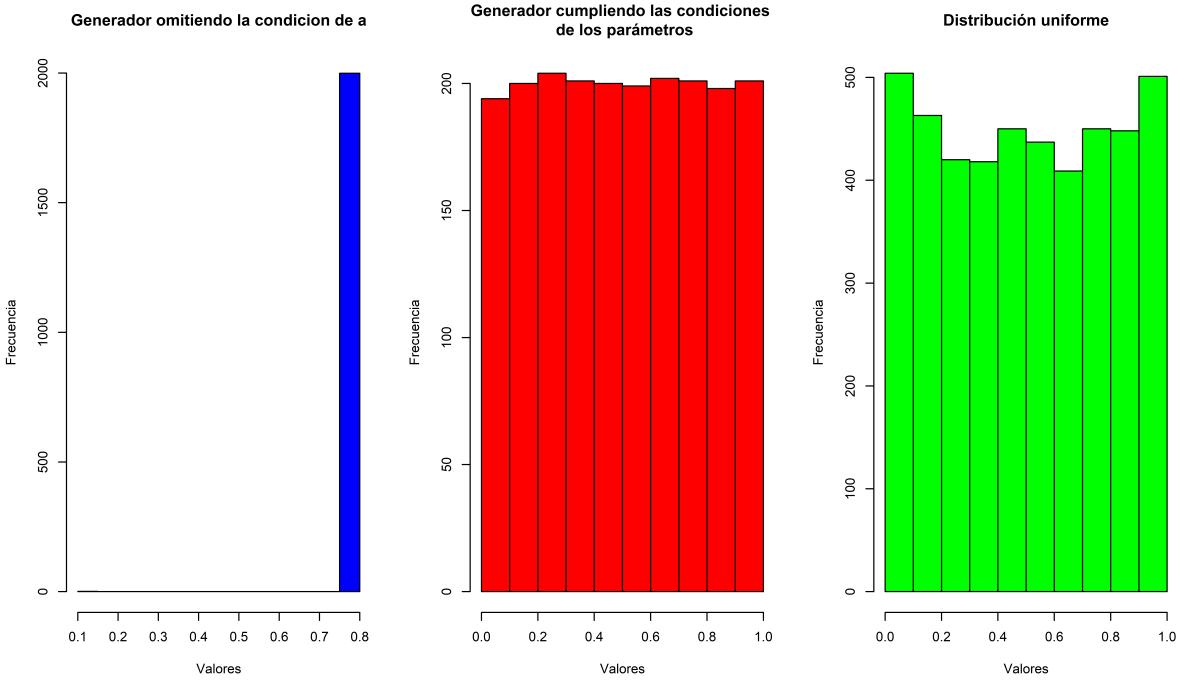


Figura 1: Comparativo de histogramas con el generador omitiendo una condición (azul), cumpliendo todas las condiciones (rojo) y el obtenido por la función `runif` (verde)

y d el número de bits que tiene una palabra de computadora en ese sistema. Para nuestro caso, $m = 2^g$, donde g es entero.

- $a = 1 + 4k$, donde k debe ser entero. Esta expresión hace que a tome valores enteros impares.
- c deber ser un entero impar y relativamente primo¹ a m .
- La semilla X_0 puede tomar cualquier valor, porque sólo afecta a la sucesión en el punto en el que comienza y no en la progresión de los números.

De acuerdo a lo anterior, se realizaron 2 experimentos, uno donde se omite uno de las condiciones y otro donde se cumplen. En ambos casos se consideró generar $n = 2000$ números pseudoaleatorios, para el primer caso, se omite la condición del parámetro a y se contemplaron los siguientes parámetros: $X_0 = 55$, $a = 8,000$, $c = 2,651$ y $m = 2,048$ y, para el segundo caso, los parámetros son: $X_0 = 70$, $a = 8,001$, $c = 2,651$ y $m = 2,048$. En la figura 1, se muestra el comparativo de los histogramas de la experimentación realizada. Como podemos observar en figura 1, al omitir una de las condiciones en el valor de los parámetros, los números generados no siguen una distribución uniforme, sin embargo, cuando sí se cumplen se puede apreciar que los números generados tienden a seguir esta distribución, por lo tanto, de no cumplirse alguna de las condiciones, el generador no será eficiente. Mediante la prueba

¹Dos números son relativamente primos si su factor común más grande es 1

`uniform.test`, también se puede apreciar que los datos del caso donde se omiten una condición, calculó un $p\text{-value} = 2,2 \times 10^{-16}$ y donde se cumplen todas las condiciones el $p\text{-value} = 1$.

Otro de los métodos para generar números pseudoaleatorios uniformes es el algoritmo congruencial cuadrático, el cual tiene la siguiente ecuación recursiva:

$$X_{i+1} = (aX_i^2 + bX_i + c) \bmod m \quad i = 0, 1, 2, \dots, N. \quad (3)$$

De igual forma, se pueden generar los números uniformes mediante la ecuación 2. García [4] menciona también que para que el generador de números pseudoaleatorios uniforme sea eficiente, se requiere que los parámetros a , b , c y m cumplan con las siguientes condiciones:

- $m = 2^g$, donde g es entero.
- a debe ser un número par.
- c debe ser un número impar
- $(b - 1) \bmod 4 = 1$.

3. Generación de números pseudoaleatorios distribuidos normalmente

En la sección anterior se mostraron diferentes algoritmos para obtener números pseudoaleatorios uniformemente distribuidos, lo que significa que todos los números tienen la misma probabilidad de aparecer en el resultado, sin embargo, hay casos en los que se hace necesario generar valores aleatorios que sigan otros tipos diferentes de distribución, como por ejemplo la distribución normal (o gaussiana).

Para la generación de los números pseudoaleatorios con distribución normal utilizaremos el método de transformación de Box-Muller [2], ya que a partir de números aleatorios uniformemente distribuidos, genera pares de números aleatorios independientes con distribución normal estándar. En el algoritmo 1 se describe el procedimiento para la generación de este par de números, donde U_1 y U_2 son variables aleatorias independientes que están uniformemente distribuidas en el intervalo $(0, 1]$, z_0 y z_1 son variables aleatorias independientes con una distribución normal con desviación típica 1.

A manera de ejemplo, se realizó una experimentación donde se generaron $n = 4,500$ réplicas con los números pseudoaleatorios distribuidos normalmente, con media ($\mu = 20$), desviación estándar ($\sigma = 15$). Los resultados obtenidos, se muestran en la figura 2, en la cual podemos observar que los histogramas de distribuciones parecen semejantes, lo que podría significar que el algoritmo utilizado al parecer está generando números pseudoaleatorios distribuidos normalmente.

Algoritmo 1: Algoritmo generador de pares de números pseudoaleatorios distribuidos normalmente (*transformación de Box-Muller*)

Entrada: media (μ), desviación estándar (σ);
Salida: Dos números pseudoaleatorios distribuidos normalmente (z_0, z_1);
 $U \leftarrow \text{runif}(2)$;
 $z_0 \leftarrow \sqrt{-2 \ln U_1} * \cos(2\pi U_2)$;
 $z_1 \leftarrow \sqrt{-2 \ln U_1} * \sin(2\pi U_2)$;
 $\text{datos} \leftarrow [z_0, z_1]$;
return $\text{datos} * \sigma + \mu$;

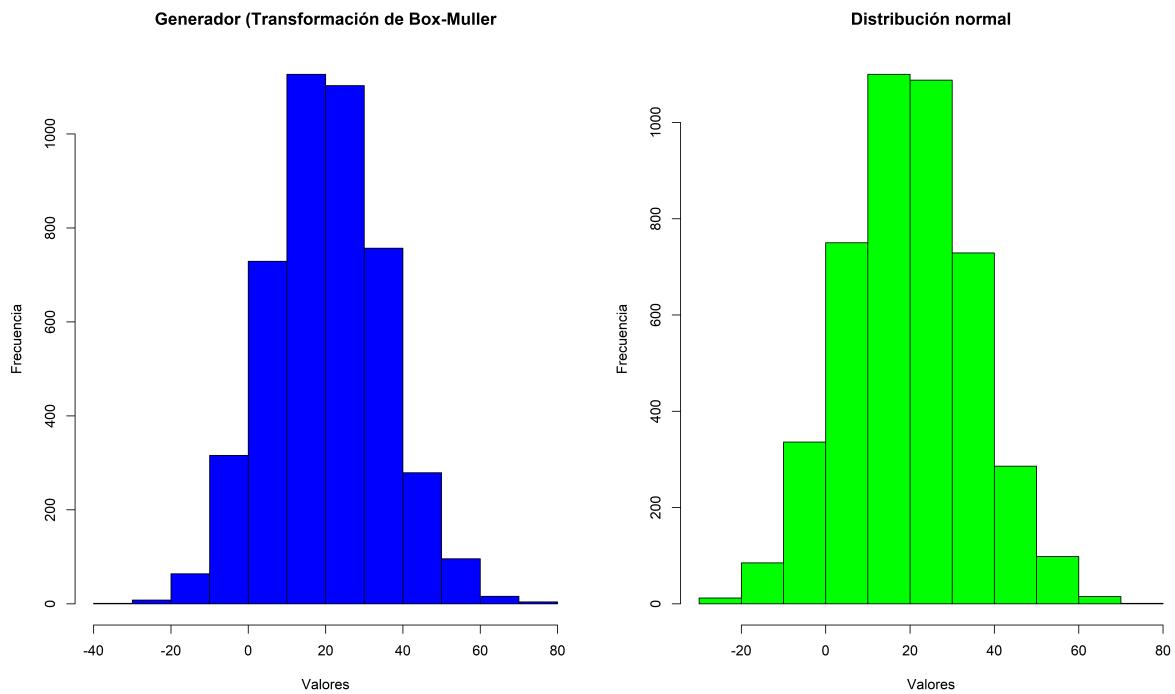


Figura 2: Comparativo de histogramas utilizando el generador propuesto (azul) y los resultados obtenidos por la función `rnorm` (verde), con $n = 4,500$ réplicas, media ($\mu = 20$) y desviación estandar ($\sigma = 15$)

Para determinar si las n réplicas de los números generados siguen una distribución normal, se aplica la prueba de normalidad *Shapiro-Wilks* con la función `shapiro.test`. Esta prueba plantea la hipótesis nula que una muestra proviene de una distribución normal y una hipótesis alternativa que sostiene que la distribución no es normal. Para aceptar la hipótesis nula el p -value debe ser mayor a 0,05. De acuerdo al resultado obtenido de la prueba, p -value = 0,7955, lo que significa que se aprueba la hipótesis nula y, por lo tanto, los datos obtenidos siguen una distribución normal.

`test.txt`

```
Shapiro-Wilk normality test

data: pseudos
W = 0.9997, p-value = 0.7955
```

3.1. Diferencias entre las variables z_0 y z_1 y comportamiento entre U_1 y U_2

Se llevaron a cabo diversos experimentos para determinar el comportamiento de los datos obtenidos por el generador propuesto, en los cuales se consideraron 3 escenarios diferentes. En el escenario 1, las variables aleatorias uniformes U_1 y U_2 no son independientes; en el escenario 2 solo se considera la variable z_0 y, en el escenario 3 solo se considera la variable z_1 . Se continua con los parámetros $n = 4,500$, $\mu = 20$, $\sigma = 15$.

Para el escenario 1, se consideran dos casos, dependencia directa ($U_2 = 2U_1$) y dependencia indirecta (recalcando la variable U_1 , siempre que esta sea menor a U_2). En la figura 3, se muestran los histogramas de ambos casos comparados con los números aleatorios generados con la función `rnorm`. En la cual podemos observar que los histogramas cuando hay dependencia directa o indirecta entre las variables aleatorias uniformes U_1 y U_2 no son similares al histograma de los números aleatorios generados por la función función `rnorm`, lo que podría indicar que cuando hay dependencia entre U_1 y U_2 , el generador propuesto, al parecer, no produce números aleatorios distribuidos normalmente. Este comportamiento se puede deber a que el generador (transformación de Box-Muller) funciona bajo el supuesto de que las variables U_1 y U_2 son números aleatorios uniformemente distribuidos, y una característica de estos números es que deben ser estadísticamente independientes.

Para hacer el análisis de los resultados de los escenarios 2 y 3, se utilizó un diagrama de cajas y bigotes incluyendo también los resultados del generador considerando las variables z_0 y z_1 y la distribución normal con la función `rnorm`. Los resultados de esta experimentación se muestran en la figura 4. En la cual podemos observar que al ejecutar el generador sólo con la variables z_0 , o con la variables z_1 o con ambas,

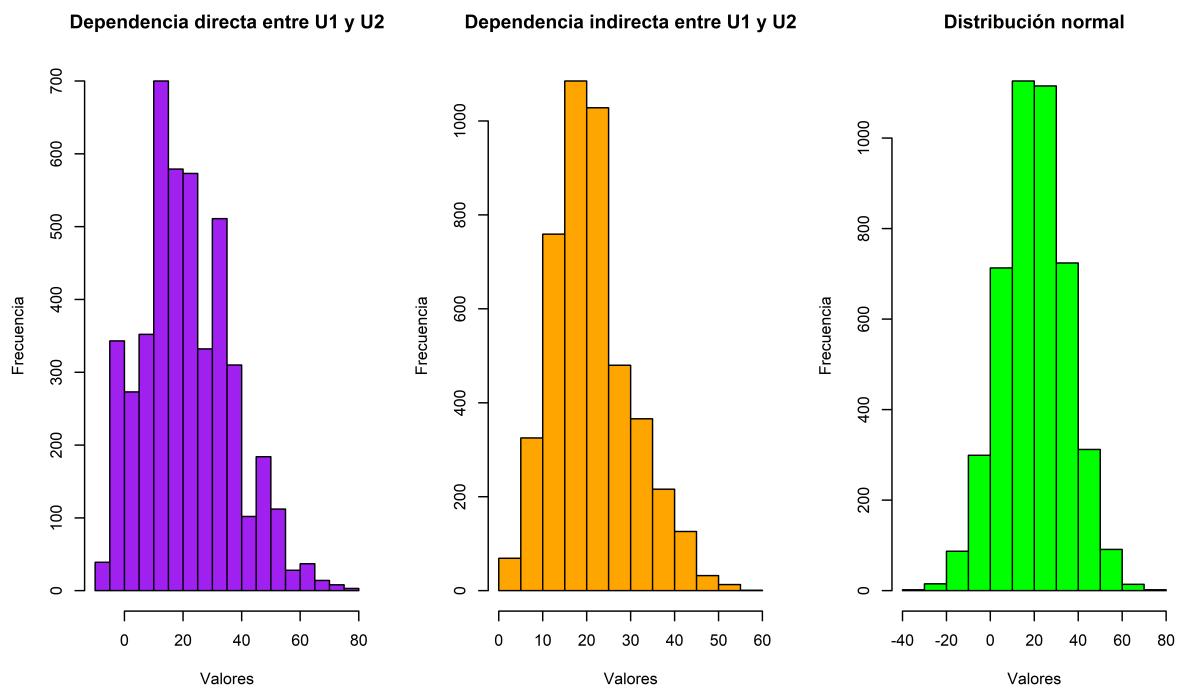


Figura 3: Comparación histogramas cuando en el generador las variables U_1 y U_2 tiene dependencia directa (morado), U_1 y U_2 tiene dependencia indirecta (naranja) y y los resultados obtenidos por la función `rnorm` (verde), con $n = 4,500$ réplicas, media ($\mu = 20$) y desviación estándar ($\sigma = 15$)

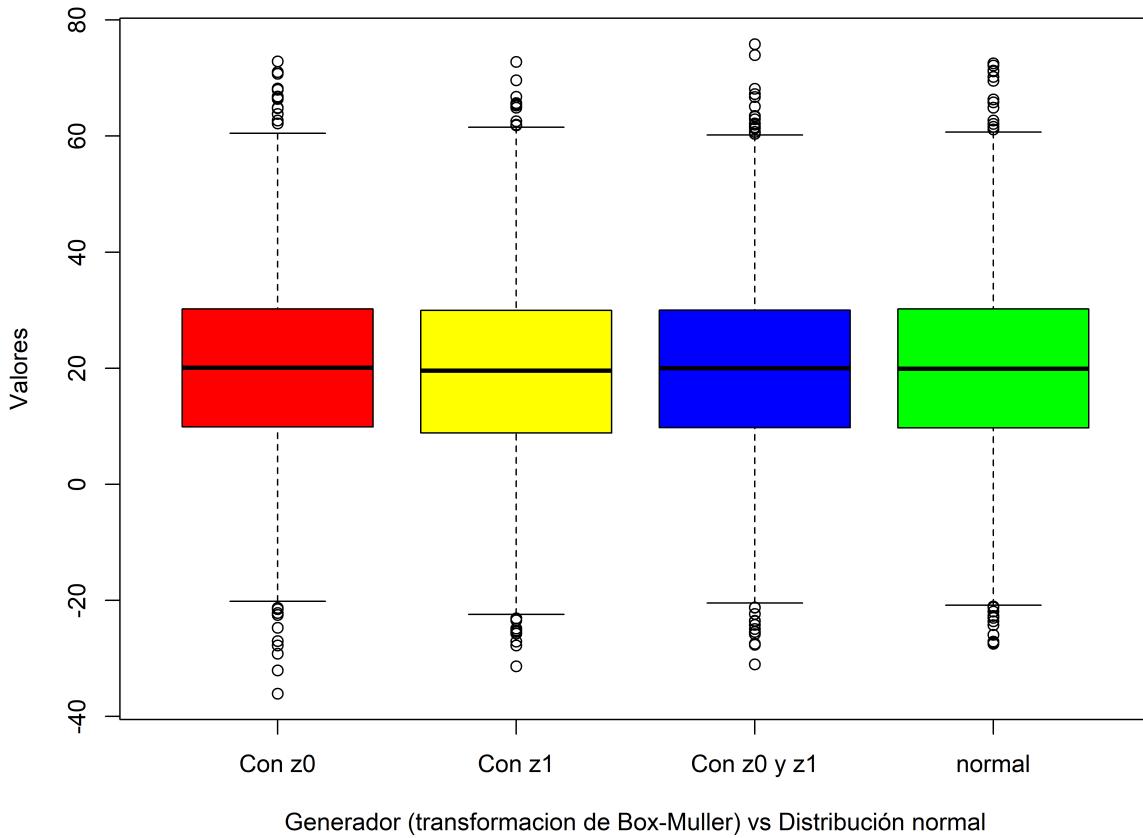


Figura 4: Comparación de diagramas de caja y bigotes del generador considerando sólo z_0 (rojo), con sólo z_1 (amarillo), ambas variables z_0 y z_1 (azul) y los resultados obtenidos por la función `rnorm` (verde), con $n = 4,500$ réplicas, media ($\mu = 20$) y desviación estándar ($\sigma = 15$)

no hay cambios significativos en los resultados obtenidos, por lo tanto, al parecer hay independencia entre estas variables, es decir, el generador puede ejecutarse solo con el cálculo de una de estas.

3.2. Uso entre generadores propuestos

A manera de práctica, se utiliza el generador de números pseudoaleatorios con distribución normal, modificando el algoritmo 1, de tal manera que las variables U_1 y U_2 se obtienen a partir del generador de números pseudoaleatorios con distribución uniforme tratado en la sección 2.

Para llevar a cabo la experimentación, se consideran los siguientes parámetros para el generador de distribución uniforme: $X_0 = 21$, $a = 8,001$, $c = 2,651$ y $m = 2,048$ con $n = 2$ réplicas y para el generador de distribución normal, $n = 4,500$ réplicas, media ($\mu = 20$) y desviación estándar ($\sigma = 15$). Los resultados

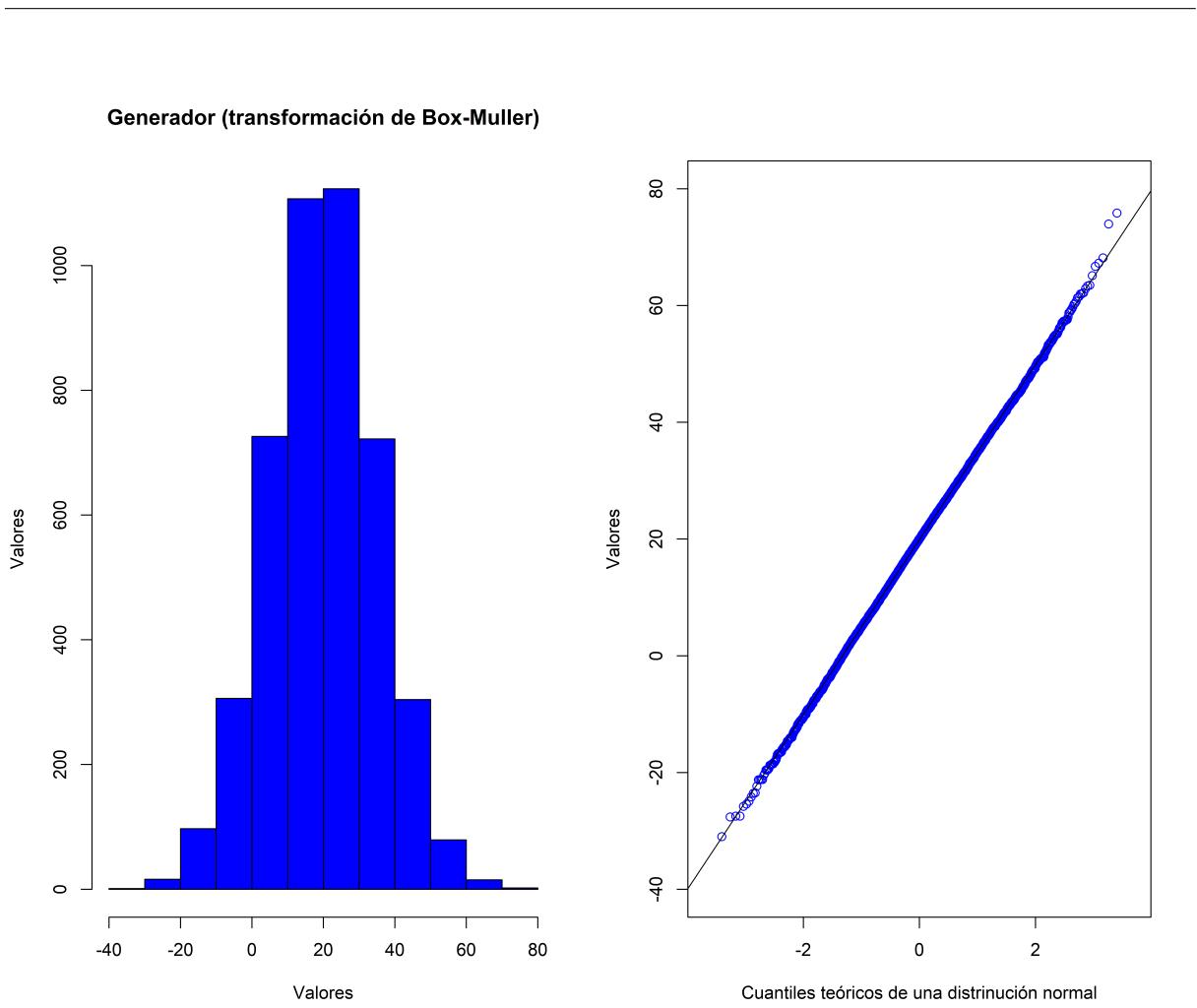


Figura 5: Correlación entre los valores obtenidos del generador (transformación de Box-Muller) y la distribución normal, con $n = 4,500$ réplicas, media ($\mu = 20$) y desviación estándar ($\sigma = 15$)

obtenidos de esta experimentación son mostrados en la figura 5. En la que podemos observar que los datos obtenidos con el generador se ajustan a una la distribución normal y, por lo tanto, el generador de números pseudoaleatorios con distribución uniforme, está arrojando valores con esta distribución.

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea5, 2020.
- [2] Box, G. E. P., Muller, Mervin E. A Note on the Generation of Random Normal Deviates. *The Annals of Mathematical Statistics*, 29(2), 1958.
- [3] Coss, Raúl Bu. *Simulación. Un enfoque práctico*. EDITORIAL LIMUSA, S.A. de C.V., 2003.

-
- [4] García D., Eduardo, García R., Heriberto, Cárdenas B., Leopoldo E. *Simulación y análisis de sistemas con Promodel*. Pearson Educación, 2006.
 - [5] Hillier, Frederick S., Lieberman, Gerald J. *Introducción a la investigación de operaciones*. McGraw-Hill, 9th edition, 2010.
 - [6] Schaeffer, Elisa. Modelos probabilistas aplicados: notas del curso. Recurso disponible en, <https://elisa.dyndns-web.com/teaching/prob/pisis/prob.htmlut2>.
 - [7] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 6

1. Pruebas estadísticas

Para realizar algunas las pruebas estadísticas se utilizó la información del Índice Nacional de Precios al Consumidor (INPC)¹ mensual por ciudades en el periodo desde enero 2015 a agosto 2020. Esta información fue consultada en la página del INEGI [3], en la sección de Precios. Se utilizó el programa R versión 4.0.2. [5] basado en la información encontrada en *Statistical Tests* [4] para la ejecución de algunas de las pruebas.

La información de las 55 ciudades del INPC y su respectivo índice con base en la segunda quincena de julio 2018, fueron descargadas en un archivo `xlsx`. Posteriormente, estos datos se guardaron en un archivo `txt` para su tratamiento en el programa R.

Para efectos del estudio, se determinaron dos conjuntos de muestras, el primer conjunto (Conjunto 1) consiste en contemplar solo la información de las 3 principales ciudades de México (Monterrey, Ciudad de México y Guadalajara) por separado. En el segundo (Conjunto 2), se consideró la información de los índices de todas las ciudades en los meses de agosto 2018, julio 2019 y agosto 2020. En todas las pruebas para aceptar la H_0 , el valor p debe ser mayor a 0.05 (nivel de significancia α).

1.1. Prueba Shapiro–Wilks

Esta prueba plantea la hipótesis nula (H_0) de que los datos provienen de una distribución normal y una hipótesis alternativa (H_1) que sostiene que la distribución no es normal.

¹El INPC es un indicador económico que muestra la variación de los precios en un periodo de tiempo.

Cuadro 1: Resultados de la prueba de Shapiro–Wilks aplicada a los Conjuntos 1 y 2

Conjunto	Contenido	Valor p
1	Monterrey	0.0002
	Cd. México	0.0002
	Guadalajara	0.0007
2	Agosto 2018	0.4888
	Julio 2019	0.0950
	Agosto 2020	0.5525

Se realizó esta prueba con la función `shapiro.test` para determinar, si los datos del Conjunto 1 y 2 siguen una distribución normal. En el cuadro 1, se muestran los valores de p obtenidos para los datos de estos conjuntos. En el cual podemos observar que en el Conjunto 1, el valor p es menor a 0,05 para cada ciudad analizada, por lo tanto, se rechaza la H_0 lo que significa que los datos analizados al parecer no siguen una distribución normal.

No obstante, para el Conjunto 2, los datos de cada mes analizado suponen seguir una distribución normal (valor $p > 0.05$), lo cual podemos observarlo también en la figura 1.

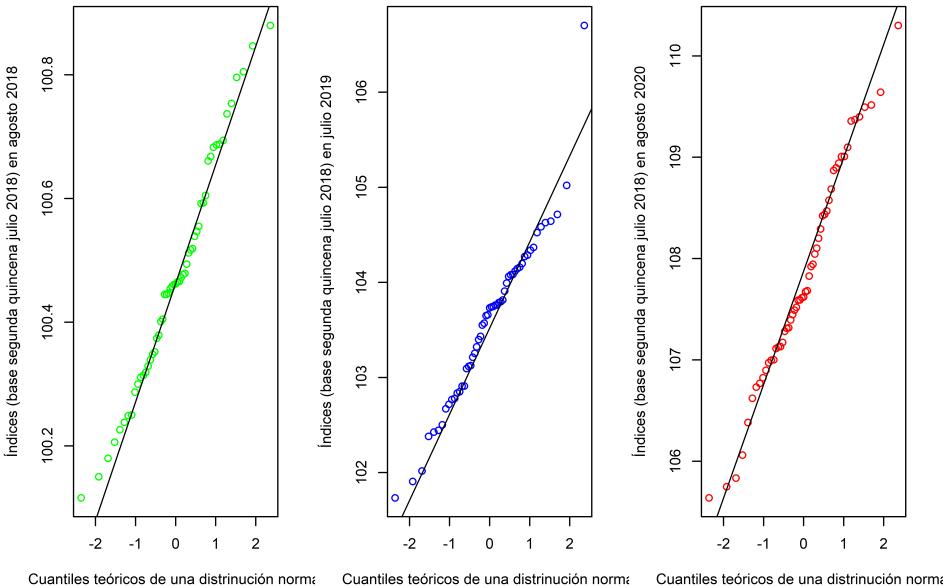


Figura 1: Comparativo de las gráficas QQ Normal de los datos con los índices de las 55 ciudades del INPC en los meses de agosto 2018 (verde), julio 2019 (azul) y agosto 2020 (rojo)

De acuerdo a lo anterior, para efectos de las pruebas que requieren datos con distribución normal, utilizaremos los datos del Conjunto 2 y para las pruebas que requieren que los datos no sigan una

distribución normal, utilizaremos los datos del Conjunto 1.

1.2. Prueba t de student

Esta prueba paramétrica se utiliza para probar si la media (μ) de una muestra con distribución normal es igual a un valor específico.

Para esta prueba se plantea como H_0 que los datos del mes de agosto 2018 tienen una $\mu = 100.4$ y, como H_1 que estos datos tienen una μ diferente. Se utilizó la función `t.test` para realizar esta prueba. De acuerdo al resultado arrojado por la función, el valor $p = 0.005324$, por lo tanto, se rechaza la H_0 , lo cual indica que los datos del mes de agosto 2018 no tienen una media $\mu = 100.4$.

1.3. Prueba de los rangos con signo de Wilcoxon

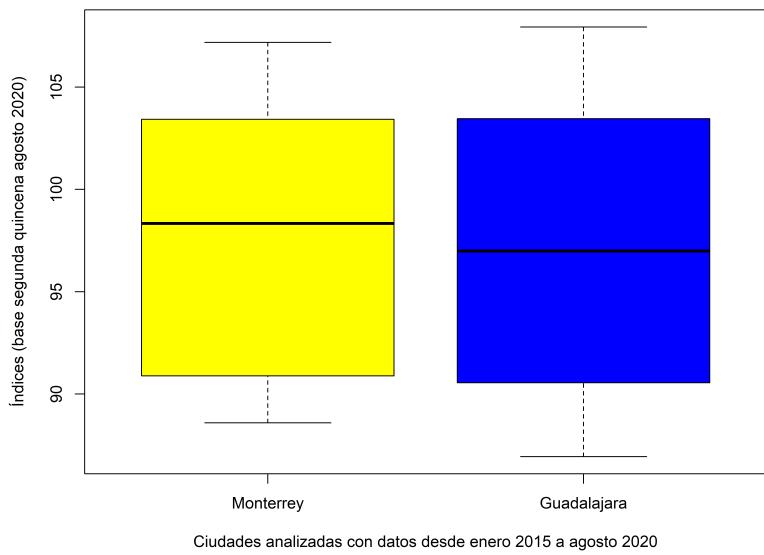
Esta prueba paramétrica se utiliza para probar si la media (μ) de una muestra, que se supone no sigue una distribución normal, es igual a un valor específico. Para esta prueba se plantea como H_0 que el promedio de los datos de la ciudad de Monterrey tienen una media de $\mu = 98.5$ y, como H_1 que en promedio estos datos no tienen esta media. Se utilizó la función `wilcox.test` para realizar esta prueba. De acuerdo al resultado arrojado por la función, el valor $p = 0.1779$, por lo tanto, no se rechaza la H_0 , lo cual indica que el promedio de los datos de la ciudad de Monterrey supone tener una $\mu = 98.5$.

1.4. Prueba t de student y Wilcoxon para dos muestras

Ambas pruebas se pueden utilizar para comparar la media de 2 muestras. La diferencia es que la prueba t supone que las muestras que se prueban siguen una distribución normal, mientras que para la prueba Wilcoxon se supone que no tienen este tipo de distribución los datos analizados.

Para la prueba de Wilcoxon se plantea como H_0 que los datos de la ciudad de Monterrey y Guadalajara tienen la misma media ($\mu_1 - \mu_2 = 0$) y, como H_1 que sus medias son diferentes ($\mu_1 - \mu_2 \neq 0$). De acuerdo al resultado arrojado por la función, el valor $p = 0.5642$, por lo tanto, no se rechaza la H_0 , lo cual indica que las medias de estas dos muestras, al parecer (una probabilidad del 56.42 %), son iguales. Esto significa que en el periodo de enero 2015 a agosto 2020, el INPC de Monterrey y Guadalajara fue similar. Este comportamiento lo podemos observar en la figura 2.

Para la prueba de t se plantea como H_0 que los datos de los meses de agosto de 2018 y agosto 2020 tienen la misma media y, como H_1 que sus medias son diferentes. De acuerdo al resultado arrojado por la función, el valor $p = 2.2 \times 10^{-16}$, por lo tanto, se rechaza la H_0 , lo cual indica que, con una probabilidad muy baja, las medias de estas dos muestras son iguales. En términos del ejercicio, de acuerdo a la figura



3, el INPC promedio de las 55 ciudades en el mes de agosto de 2018 ($\mu_1 = 100.4723$) fue más bajo que el promedio reportado en agosto de 2020 ($\mu_2 = 107.8254$), es decir, son más costosos los precios de la canasta de bienes y servicios en el mes agosto 2020 que hace 2 años.

1.5. Prueba de Kolmogorov–Smirnov

Esta prueba se utiliza para comprobar si 2 muestras siguen la misma distribución. Para su desarrollo se plantea como H_0 que los datos de Cd. México y Guadalajara tienen una misma distribución y, como H_1 que estas muestras no tienen la misma distribución. Se utilizó la función `ks.test` para realizar esta prueba. De acuerdo al resultado arrojado por la función, el valor $p = 0.7384$, por lo tanto, no se rechaza la H_0 , lo cual indica que, al parecer, los datos de Cd. México y Guadalajara tienen una misma distribución.

test.txt

```

Two-sample Kolmogorov-Smirnov test

data: cdMexico and guadalajara
D = 0.11765, p-value = 0.7384
alternative hypothesis: two-sided

```

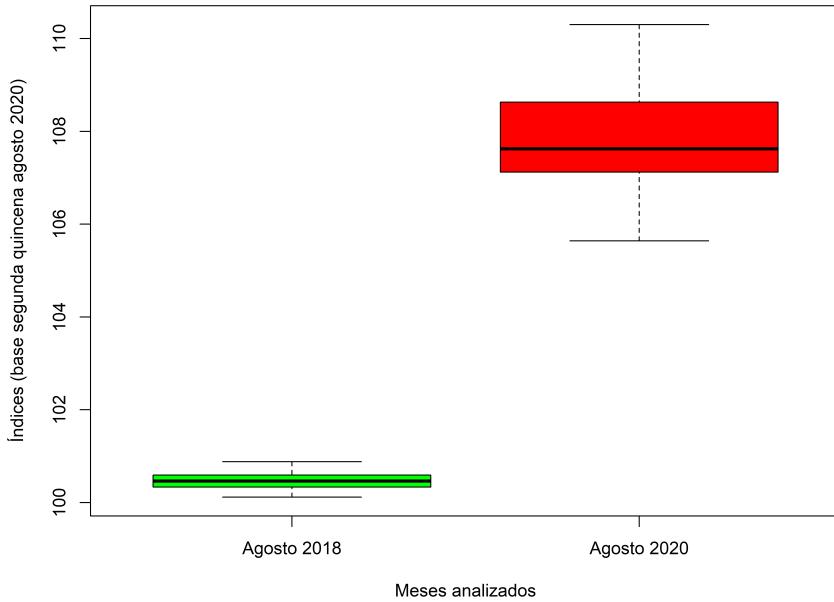


Figura 3: Comparativo de los diagramas de caja y bigotes con los datos recopilados de las 55 ciudades del INPC en los meses agosto 2018 (verde) y agosto 2020 (rojo)

1.6. Prueba F de Fisher

Es una prueba paramétrica que se utiliza para verificar si dos muestras tienen la misma varianza. Se plantea como hipótesis nula (H_0) que los datos de los meses de agosto 2018 con julio 2019 tienen la misma varianza y, como hipótesis alternativa (H_1) que estas muestras sus varianzas son diferentes. Se utilizó la función `var.test` para realizar esta prueba. De acuerdo al resultado obtenido por la función, el valor $p = 2.2 \times 10^{-16}$, por lo tanto, se rechaza la H_0 , lo cual indica que los datos de los meses de agosto 2018 y julio 2019 no tienen la misma varianza.

test.txt

```

F test to compare two variances

data: agosto2018 and julio2019
F = 0.044063, num df = 54, denom df = 54, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.02570083 0.07554563
sample estimates:
ratio of variances
 0.04406342

```

Cuadro 2: Tabla de contingencia para la prueba Chi²

	Monterrey	Cd. México
Ene 2016	90.779	88.461
Feb 2016	91.292	88.878
Mar 2016	91.400	89.077
Abr 2016	90.261	88.921
May 2016	90.250	88.946
Jun 2016	90.314	88.961
Jul 2016	90.505	89.381
Ago 2016	90.992	89.636
Sep 2016	91.635	90.216
Oct 2016	92.835	90.383
Nov 2016	93.108	90.620
Dic 2016	93.460	91.052

Esta prueba también se puede usar para saber si la varianza de una muestra sigue un determinado valor, solo hay que adicionar ciertos parámetros en la función `var.test` para realizarla. Se pueden ver ejemplos en el repositorio de Freddy [2].

1.7. Prueba Chi²

Esta prueba se puede utilizar para probar si dos variables categóricas son dependientes, mediante una tabla de contingencia. Para realizar esta tabla, se consideraron dos factores, ciudades y meses del año. Para las ciudades, se contemplan Monterrey y Guadalajara y se analizaran los índices de los 12 meses del año 2016. En el cuadro 2, se encuentra la información de los dato a analizar.

Se plantea como H_0 que valores del índice alcanzado de cada ciudad es independiente del mes del año en que se encuentren y, como H_0 las variables son dependientes. Se utilizó la función `chisq.test` para realizar esta prueba. Para aceptar la H_0 se deben de cumplir dos consideraciones, que el valor $p > 0.05$ y que el $X - Square$ sea menor al valor crítico. Para calcular este valor se utiliza la función `qchisq(0.95, n-1)`, donde 0.95 es el nivel de confianza y n-1 son los grados de libertad. Estos grados de libertad dependen de la cantidad de filas y columnas que tenga la tabla de contingencia y se calculan con la siguiente formula: $(filas - 1)*(columnas - 1)$. Para este caso, el valor crítico = 19.67514.

Al aplicar la función se obtiene que el valor $p = 1$ y un $X - Square <$ valor crítico, por lo tanto, no hay suficiente evidencia estadística para rechazar H_0 , lo cual indica que las variables son independientes, es decir, el índice alcanzado de las ciudades de Monterrey y Guadalajara no dependen del mes del año.

test.txt

```
Pearson's Chi-squared test

data: tablacontingencia
X-squared = 0.035179, df = 22, p-value = 1
```

1.8. Prueba de correlación

Es una prueba paramétrica que se utiliza para probar si hay una relación lineal de dos variables continuas. Se pretende determinar si hay una correlación entre los índices de los de meses de julio 2019 y agosto 2020. Se plantea entonces como H_0 que los datos del mes de julio 2019 no están relacionados con el mes de agosto 2020 y, como H_1 existe una relación entre estas variables. Se emplea la función `cor.test`. De acuerdo a los resultados arrojados por la prueba, el valor $p = 1.136 \times 10^{-9}$, por lo tanto, se rechaza la H_0 , lo cual indica que, los datos de los meses de julio 2019 y agosto 2020 tienen alguna correlación. En la figura 4 se muestra el diagrama de dispersión de los datos en el que visualmente se puede observar esta correlación.

test.txt

```
Pearson's product-moment correlation

data: julio2019 and agosto2020
t = 7.3722, df = 53, p-value = 1.136e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.5500801 0.8217206
sample estimates:
cor
0.711539
```

2. Preguntas

Relación entre contraste de hipótesis y pruebas estadísticas

Las pruebas estadísticas son técnicas que utilizan muestras representativas de una población para evaluar la evidencia que proporcionan los datos y la hipótesis es la suposición de algún fenómeno o problema la cual son comprobadas a través de las pruebas estadísticas. En estas pruebas existen dos tipos de hipótesis, nula (H_0) y la alternativa (H_1 o H_A).

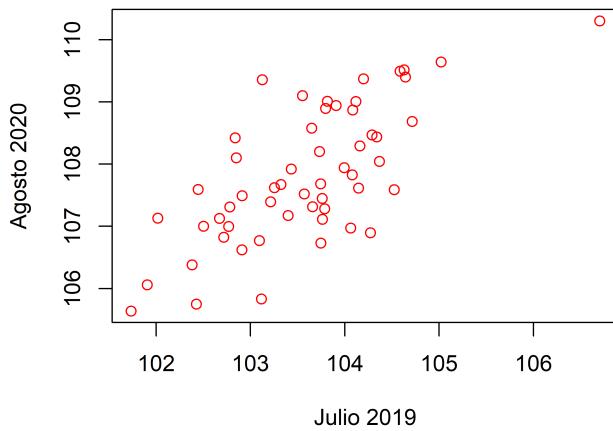


Figura 4: Diagramas de dispersión de los índices en los meses de julio 2019 y agosto 2020

¿Qué indicaría rechazar la hipótesis nula?

Normalmente, la hipótesis nula (H_0) establece la igualdad entre las medias, varianzas, entre otros, por lo tanto, si se rechaza indicaría que existe alguna diferencia entre ellas.

¿Cómo se interpreta la salida de una prueba estadística?

Al diseñar un estudio, se especifica un nivel de significación alfa (α) que debería estar entre 0 y 1, lo que indica que por encima este valor H_0 no debería ser rechazada. La prueba estadística produce un número denominado valor p el cual también se encuentra entre 0 y 1. En términos más prácticos, el valor p se compara con α , si $p < \alpha$, rechazamos H_0 y aceptamos H_a con un riesgo proporcional al valor p de ser errónea. Por otro lado, si $p > \alpha$, no rechazamos H_0 , pero esto no implica necesariamente que debamos aceptarla, solo que nuestro experimento y nuestra prueba estadística no han sido suficientemente “fuertes” para producir un valor p inferior a α .

¿Cómo seleccionar el alpha?

El nivel de significancia se denota con la letra griega alfa α . No existe una evidencia científica de cuál sea el valor más adecuado. Sin embargo, el valor más empleado es 0.05, se suelen tomar valores pequeños (menores al 10 %), esto debido a que representa el riesgo de rechazar la hipótesis nula (H_0) cuando es verdadera. En ese sentido con una buena elección de α se delimita muy bien cuando rechazar la H_0 lo que aumenta la probabilidad de tomar la decisión correcta.

¿Cuáles son los errores frecuentes de interpretación del valor p ?

La interpretación del valor p conduce al rechazo o la aceptación de la hipótesis nula, específicamente

Cuadro 3: Situaciones posibles al probar una hipótesis estadística

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

el valor p corresponde al menor valor de alfa (α) que ocasiona el rechazo de la hipótesis nula (H_0). Los errores más frecuentes que se presentan son los errores tipo I y tipo II. El error tipo I se presenta cuando se rechaza la H_0 y es verdadera, mientras que el error tipo II consiste en no rechazar la H_0 cuando es falsa. De acuerdo a Walpole [6], en el cuadro 3 se resumen estos errores. La probabilidad de cometer esos errores se reduce aumentando el tamaño de la muestra.

¿Qué es la potencia estadística y para qué sirve?

La potencia estadística o el poder del estadístico, corresponde a la capacidad que tiene una prueba de llevar al rechazo de la hipótesis nula (H_0). La potencia estadística aumenta con el valor de alfa α , con la precisión de las medidas y el número de repeticiones, también depende del tipo de prueba estadística que se esté realizando. Este parámetro puede ser calculado antes o después de realizar el experimento.

Ejemplos de pruebas estadísticas paramétricas y no paramétricas

Las pruebas paramétricas se emplean para datos numéricos, suelen estar basadas en las propiedades de la distribución normal para la variable dependiente. Es decir, los datos son mediciones repetidas de la misma variable, muestreado de la población realizado al azar y cuando la muestra es grande. Como ejemplo de pruebas paramétricas se tienen:

- La “t” de student.
- El coeficiente de correlación de Pearson.
- La regresión lineal.
- Análisis de varianza unidireccional (ANOVA *Oneway*).
- Análisis de varianza factorial (ANOVA).
- Análisis de covarianza (ANCOVA).
- Estadígrafos descriptivos como la desviación estándar, la moda, la mediana y la media.

Por otra parte, las pruebas no paramétricas se aplican con variables nominales y ordinales, no asume un tipo específico de distribución. Ejemplos de este tipo de pruebas son:

-
- La X^2
 - Coeficientes de correlación e independencia para tabulaciones cruzadas.
 - Coeficientes de correlación por rangos ordenados Spearman y Kendll.

Resume LA GUIA para encontrar la prueba estadística que buscas

- Escribir claramente el objetivo de análisis.
- Tipo de variables.
- Si son muestras independientes o no.
- Identificar si se pueden aplicar técnicas paramétricas.
- Seleccionar la prueba adecuada.
- Realizar la prueba de hipótesis.
- Interpretar y graficar los resultados.

En la figura 5, se describe el proceso de selección de una prueba estadística basado en el estudio realizado por Florez-Ruiz [1].

¿Cuáles son los supuestos para aplicar técnicas paramétricas?

Las pruebas paramétricas están basadas en la distribución normal para la variable dependiente, y los requisitos para aplicarlas son las siguientes:

- Las observaciones deben ser independientes entre sí.
- Las poblaciones deben hacerse en poblaciones distribuidas normalmente.
- Estas poblaciones deben tener la misma varianza.
- Las variables deben haberse medido por lo menos en una escala de intervalo de manera que sea posible utilizar las operaciones aritméticas.

Referencias

- [1] Flores-Ruiz, E., Miranda-Novales, María, Villasis-Keever, Miguel. The research protocol VI: How to choose the appropriate statistical test. Inferential statistics. *Revista Alergia Mexico*, 64:364–370, 07 2017.

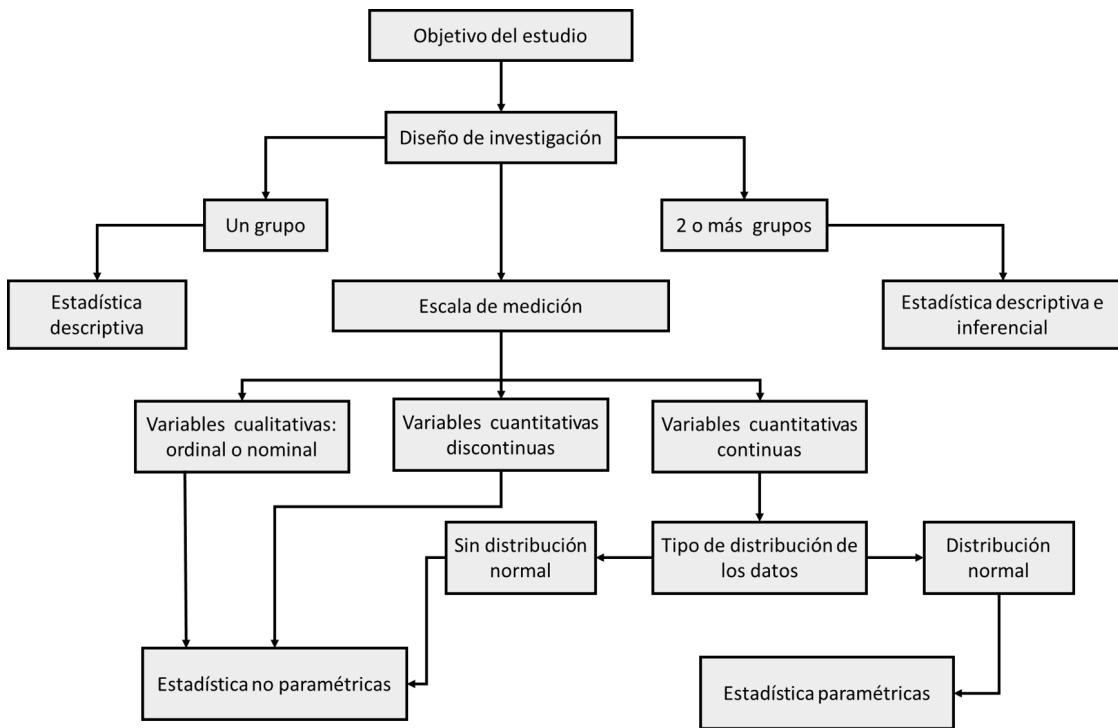


Figura 5: Proceso de selección de una prueba estadística

- [2] Hernandez, Freddy. Prueba de hipótesis.
- [3] INEGI. Índice Nacional de Precios al Consumidor (INPC). Recurso libre, disponible en <https://www.inegi.org.mx/datos/>, 2020.
- [4] Prabhakaran, Selva. Statistical Tests. Recurso disponible en: <http://r-statistics.co/Statistical-Tests-in-R.html>, 2017.
- [5] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [6] Walpole, Ronald E., Myers, Raymond H., Myers Sharon L., Ye, Keying. *Probability statistics for Engineers Scientists*. Pearson Education, Inc., 9th edition, 2012.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 7

1. Introducción

En el presente trabajo se analizan datos generados con diferentes funciones para mostrar la aplicación de la regresión lineal y el uso de las transformaciones. Este análisis es complementado con resultados obtenidos de experimentos computacionales realizados en el software R versión 4.0.2 [4]. El código empleado se encuentra en el repositorio GitHub [1].

2. Regresión lineal

El análisis de regresión se centra en la exploración, explicación y estudio de dependencia de una variable mediante una o más variables explicativas. El término regresión significa que utilizaremos información pasada y es lineal porque está bajo el supuesto de que entre dos variables (x y y) existe una relación lineal. Cuando se utiliza sólo una variable independiente para tratar de explicar la variable dependiente, es una regresión lineal simple, pero cuando se utilizan más de dos variables independientes o dependientes, se conoce como regresión múltiple.

La regresión lineal consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependientes o respuesta se le identifican como y , y a la variables predictoras o independientes como x . Las ecuaciones 1 y 2, son las ecuaciones estimadas de regresión lineal simple y múltiple, respectivamente,

$$y = b_1x + b_0 + \epsilon, \quad (1)$$

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \epsilon, \quad (2)$$

donde b_0 es la ordenada al origen (el valor de y cuando x es igual a cero) y, b_1 es la pendiente de la recta (el cambio en y cuando x aumenta en una unidad) y ϵ es el error aleatorio. Este último representa la diferencia entre el valor ajustado por la recta y el valor real. Recoge el efecto de todas aquellas variables que influyen en y pero que no se incluyen en el modelo como predictores. Al error aleatorio también se le conoce como residuo. Los errores, se consideran variables aleatorias independientes distribuidas normalmente con media cero y desviación estándar σ . Esto implica que el valor medio o valor esperado de \hat{y} son los mostrados en las ecuaciones 3 y 4, las cuales son las ecuaciones estimadas de regresión lineal simple y múltiple, respectivamente,

$$\hat{y} = \hat{b}_1 x + \hat{b}_0 + \epsilon, \quad (3)$$

$$\hat{y} = \hat{b}_1 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p + \epsilon, \quad (4)$$

donde \hat{y} es el valor estimado (aproximado) de y , \hat{b}_0 y \hat{b}_p son estimaciones que se conocen como coeficientes de regresión, ya que toman aquellos valores que minimizan la suma de cuadrados residuales, dando lugar a la recta que pasa más cerca de todos los puntos.

En conclusión, el análisis de regresión consiste en definir la variable independiente x que ayude a explicar (estimar) la variable dependiente y , siempre que exista una relación lineal entre ellas, además, de que ambas variables deben ser cuantitativas [2].

Para determinar si hay relación lineal entre las variables dependientes e independientes, por lo general, se utilizan las gráficas de dispersión, ya que son ayudas visuales que permiten observar, rápidamente, si existe esta relación. Sin embargo, un análisis más fuerte es determinar la correlación entre los datos, ya que es una medida de la presencia de una relación lineal en datos bivariados (entre dos variables). Hay diferentes definiciones, sin embargo, la comúnmente utilizada, es la de correlación de Pearson (se puede hallar con la función `cor(x, y)`). Se denota con la letra r y sus valores se interpretan de la siguiente manera:

- Cerca de uno: cuando x crece, y crece de manera linealmente dependiente.
- Cerca de menos uno: cuando x crece, y disminuye de manera linealmente dependiente (o vice versa).
- Cerca de cero: no está presente ninguna relación lineal entre x y y .

El valor de la correlación se puede hallar con la función `cor(x, y)` en el programa R. De igual manera, la función `lm(y~x)`, donde y , es la variable dependiente (la que se trata de predecir) y x es la variable predictora (independiente), permite obtener el modelo estimado de regresión lineal de los datos a analizar. Con la función de `summary` se generan los resultados de este modelo.

Para entender el funcionamiento de la función `lm` se realizó una serie de experimentos, en los cuales consideramos crear diversas funciones, tanto lineales como no lineales, para la obtención de datos dependientes (variables y).

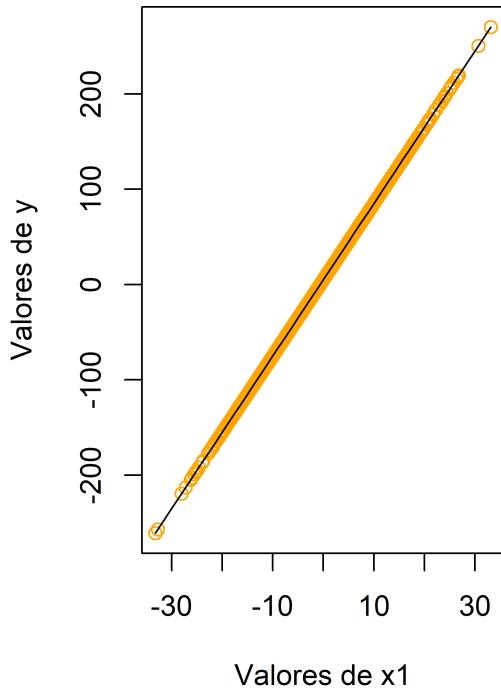


Figura 1: Diagrama de dispersión de los datos provenientes de la ecuación lineal

Para el ejemplo de la regresión lineal simple, se generaron 1,000 datos a partir de la ecuación $y = 8x_1 + 5$, donde x_1 es un número pseudoaleatorio entre 10 y 60. Los resultados obtenidos fueron guardados en un dataframe. Se realizó un diagrama de dispersión (ver figura 7 para apreciar la relación entre las variables x_1 y y , en la cual se observa, claramente, la relación lineal entre estas variables.

test.txt

```

Call:
lm(formula = y ~ x1, data = f1)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.870e-14 -2.100e-14 -1.760e-14 -1.360e-14  1.241e-11 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.000e+00  3.532e-14 1.415e+14 <2e-16 ***
x1          8.000e+00  9.188e-16 8.707e+15 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```

```

Residual standard error: 4.281e-13 on 998 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.58e+31 on 1 and 998 DF,  p-value: < 2.2e-16

```

De acuerdo al resultado obtenido de la aplicación de la función `lm`, se observa que los coeficientes de regresión $\hat{b}_1 = 8$ y $\hat{b}_0 = 5$ y $r = 1$, es decir, cuando $x_1 = 0$, es valor estimado de $y = 8$ y por cada unidad que aumente x_1 el valor estimado de y aumenta en 8 unidades. Por lo tanto, el modelo estimado de regresión lineal para estos datos estaría dada como $\hat{y} = 8x_1 + 5$ (el cual con anterioridad se tenía). También se realizó el mismo procedimiento considerando que los valores de la variable x_1 tuvieran una distribución exponencial, normal o uniforme, y se obtuvieron los mismos resultados.

Para la regresión lineal múltiple, también se generaron 1,000 datos a partir de la ecuación $y = 20 * x_1 + 5 * x_2 + 1$, donde las variables x_1 y x_2 son pseudoaleatorias uniformes. Los resultados obtenidos fueron guardados en un data frame. Se utilizó la función `cor` para determinar la correlación entre estas variables. Las respectivas correlaciones se encuentran graficadas en la figura 2, en la cual se puede observar la dependencia lineal de la variable y con las variables x_1 y x_2 .

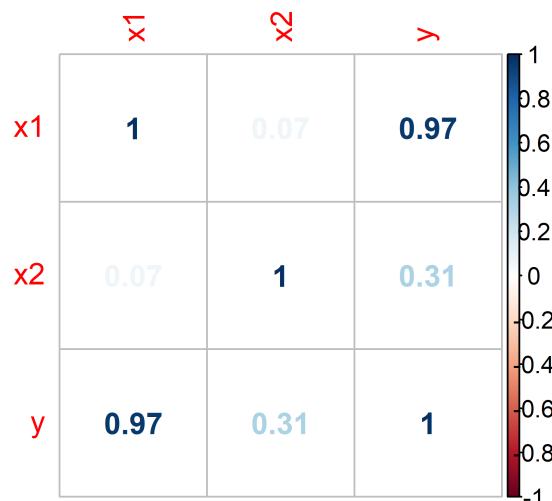


Figura 2: Correlación entre los datos provenientes de una ecuación lineal con dos variables independientes

```

test.txt
Call:
lm(formula = y ~ x1 + x2, data = f3)

Residuals:

```

```

Min           1Q       Median        3Q       Max
-3.578e-13 -2.000e-16  3.200e-16  8.600e-16  2.560e-14

Coefficients:
            Estimate Std. Error   t value Pr(>|t|)
(Intercept) 1.000e+00 9.223e-16 1.084e+15 <2e-16 ***
x1          2.000e+01 1.249e-15 1.601e+16 <2e-16 ***
x2          5.000e+00 1.270e-15 3.938e+15 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.139e-14 on 997 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
F-statistic: 1.401e+32 on 2 and 997 DF,  p-value: < 2.2e-16

```

De acuerdo al resultado obtenido de la aplicación de la función `lm`, se observa que $\hat{b}_1 = 20$, $\hat{b}_2 = 5$, $\hat{b}_0 = 1$ y $r = 1$. Por lo tanto, el modelo estimado de regresión lineal múltiple para estos datos estaría dada como $\hat{y} = 20x_1 + 5x_2 + 1$. También se realizó el mismo procedimiento considerando que los valores de las variables x_1 y x_2 tuvieran una distribución exponencial, normal o sin una distribución específica y se obtuvieron los mismos resultados.

Hasta aquí, se ha realizado el análisis con datos de los cuales se conocía su comportamiento, sin embargo, en los casos reales, este comportamiento no se conoce y no siempre es lineal, lo cual aumenta la complejidad del análisis de estos y la predicción de su comportamiento, haciendo necesario ajustar la tendencia de los datos a un modelo lineal ya que esto facilita el análisis y la predicción de su comportamiento para la toma de decisiones. Para llevar a cabo este ajuste se utilizan las transformaciones.

3. Transformaciones

La meta de las transformaciones es acomodar los datos a una relación lineal, ya que bajo este criterio se facilita el análisis y la predicción del comportamiento de estos.

De acuerdo a Mangiafico [3], para datos sesgados a la derecha o izquierda (sesgo positivo y negativo, respectivamente), las transformaciones comunes incluyen raíz cuadrada, raíz cúbica y logaritmo. Otro enfoque es utilizar la transformación de Box-Cox, la cual determina un valor lambda (λ), que se utiliza como coeficiente de potencia para transformar los valores. El procedimiento Box-Cox se realiza mediante la función `boxcox` en el programa R. Esta utiliza un procedimiento de *log-likelihood* para encontrar la lambda que se utilizará para transformar la variable dependiente de un modelo lineal (una regresión lineal). Cabe mencionar que estas transformaciones solo se pueden aplicar sobre variables de respuesta

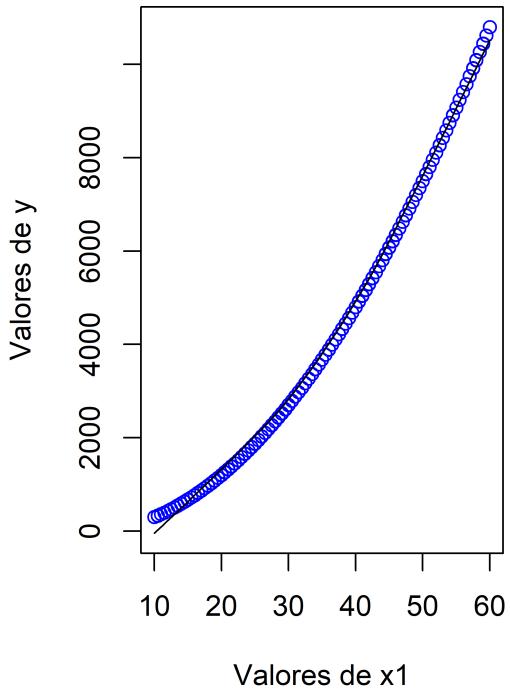


Figura 3: Diagrama de dispersión de los datos generados a partir de una función cuadrática

(variables dependientes) estrictamente positivas.

En este trabajo se utilizó las transformaciones logarítmicas, raíz cuadrada y Box–Cox. Para determinar que tipo de transformación es la más apropiada, se utilizó el coeficiente de determinación (R^2), el cual es una medida de la bondad de ajuste de la recta estimada a los datos reales, es decir, entre más cercano esté este valor a la unidad, la transformación tendrá un mejor ajuste. Este coeficiente se calcula elevando al cuadrado el valor de r en los modelos de regresión lineal simple el valor de (R^2), sin embargo, no es así en regresión múltiple. Existe una modificación de (R^2) conocida como ($R^2 - \text{ajustado}$) que se emplea principalmente en los modelos de regresión múltiple, el cual introduce una penalización cuantos más predictores se incorporan al modelo.

Para el análisis de las trasformaciones, se generaron 1,000 datos a partir de la ecuación cuadrática de la forma $y = 3x_1^2 + rnorm(n)$, donde x_1 es un número pseudoaleatorio entre 10 y 60 y n son las réplicas. Los resultados obtenidos fueron guardados en un data frame. Se realizó un diagrama de dispersión (ver figura 3 para apreciar la relación entre las variables x_1 y y), en la cual se observa, claramente, que hay una relación entre estas variables, pero no es lineal, por lo que se procedió a realizar tres diferentes transformaciones.

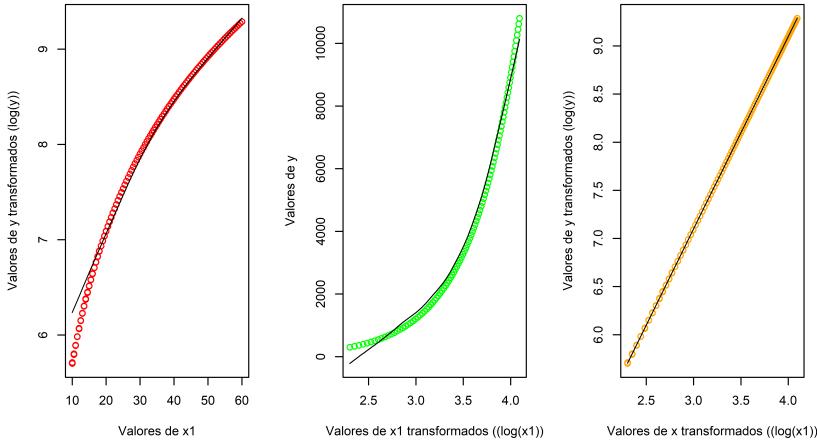


Figura 4: Comparativo de diagramas de dispersión con la transformación logarítmica

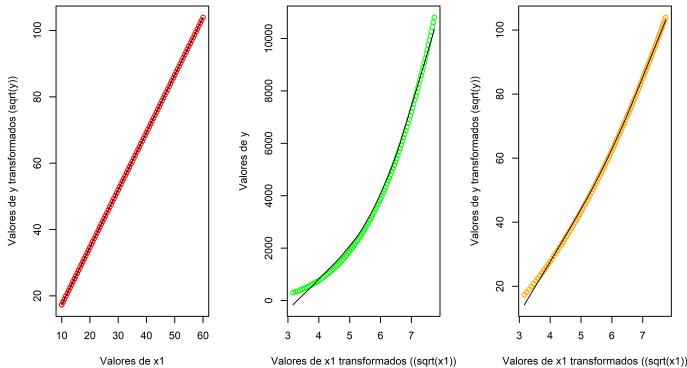


Figura 5: Comparativo de diagramas de dispersión con la transformación de raíz cuadrada

En la figura 4, se muestra el comparativo de los diagramas de dispersión de las transformaciones logarítmicas aplicadas, tanto a los datos de las variables x , como a y y a ambas, en la cual podemos observar que se obtiene un ajuste lineal a los datos cuando se aplica la transformación logarítmica a x y y .

En la figura 5, se muestra el comparativo de los diagramas de dispersión de las transformaciones de raíz cuadrada aplicadas, tanto a los datos de las variables x , como a y y a ambas, en la cual podemos observar que se obtiene un ajuste lineal a los datos cuando se aplica la transformación logarítmica a los datos de la variable, así mismo, su coeficiente de determinación $R^2 = 1$.

Finalmente, en la figura 6, se muestra el diagrama de dispersión de la transformación Box–Cox, en la cual se puede observar que al igual que con la transformación logarítmica de x y y o la raíz cuadrada de y , los datos se logran ajustar a relación lineal. Además, para estas trasformaciones el coeficiente de

determinación fue $R^2 = 1$. El valor de lambda para estos datos es de $\lambda = 0.505$.

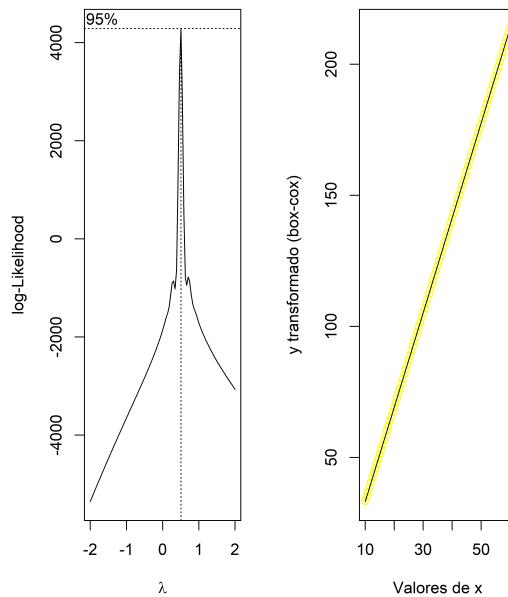


Figura 6: Valores de lambda (λ) y diagrama de dispersión con la transformación de Box–Cox

test.txt

```

Call:
lm(formula = (y^lambdabox - 1)/lambdabox ~ x1, data = f2)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.16687 -0.09227 -0.02531  0.07517  0.44237 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.074042   0.008967 -342.8   <2e-16 ***
x1          3.609200   0.000232 15559.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1065 on 998 degrees of freedom
Multiple R-squared:      1,        Adjusted R-squared:      1 
F-statistic: 2.421e+08 on 1 and 998 DF,  p-value: < 2.2e-16

```

Para efectos prácticos, se consideraron los datos obtenidos de la aplicación de la función `lm` con la transformación de Box–Cox (aunque también es válido escoger cualquiera de las tres buenas trasformaciones obtenidas), para determinar el modelo estimado de regresión lineal simple. De acuerdo a los

resultados arrojados, se observa que $\hat{b}_1 = 3.60$, $\hat{b}_0 = -3.07$. Por lo tanto, el modelo estimado de regresión lineal simple para estos datos estaría dado como $\hat{y} = 3.60x_1 - 3.07$. Adicionalmente, se realizó el mismo procedimiento considerando que los valores de la variable x_1 tuvieran otro tipo de distribución y para este caso, no se obtuvieron los mismos resultados, haciendo que las transformaciones propuestas no fueron suficientes. En el programa R, la rutina `assumptions(lm)` de la librería `trafo`, permite obtener información acerca de varias transformaciones que cumplan con los supuestos de un modelo lineal de los datos que estamos analizando.

Para la regresión lineal múltiple, también se generaron 1,000 datos a partir de la ecuación no lineal con dos variables independientes. Los resultados obtenidos fueron guardados en un dataframe. Se utilizó la función `cor` para determinar la correlación entre estas variables. Las respectivas correlaciones se encuentran graficadas en la figura 2, en la cual se puede observar que no hay una dependencia lineal de y con las variables x_1 y x_2

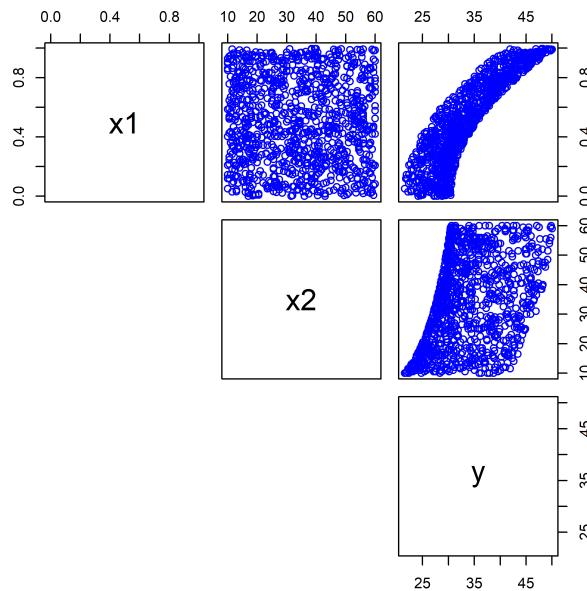


Figura 7: Correlación entre los datos provenientes de una ecuación no lineal con dos variables independientes

Se aplicaron las transformaciones logarítmicas, de raíz cuadrada y Box–Cox, de las cuales, la transformación ($\log y$) presentó el mejor $R^2 = 0.9519$, es decir, es capaz de explicar el 95.19 % de la variabilidad observada en los datos. De acuerdo a los resultados arrojados, se observa que $\hat{b}_1 = 0.586$, $\hat{b}_2 = 0.005$ y $\hat{b}_0 = 3.035$. Por lo tanto, el modelo estimado de regresión lineal múltiple para estos datos estaría dado como $\hat{y} = 0.586x_1 + 0.005x_2 + 3.035$, lo que significa que, si el resto de variables se mantienen constantes, por cada unidad que aumenta el predictor en cuestión, la variable y varía en promedio tantas unidades como indica la pendiente. Para este ejemplo, por cada unidad que aumenta el predictor x_1 , el

valor de y aumenta en promedio 0.586 unidades, manteniéndose constante la variable x_2 .

test.txt

```
Call:
lm(formula = log(y) ~ x1 + x2, data = f4)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.11258 -0.02824 -0.00467  0.02808  0.11054 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.035e+00  4.049e-03  749.50   <2e-16 ***
x1          5.863e-01  4.517e-03  129.81   <2e-16 ***
x2          5.089e-03  8.849e-05   57.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04128 on 997 degrees of freedom
Multiple R-squared:  0.9519,    Adjusted R-squared:  0.9518 
F-statistic:  9861 on 2 and 997 DF,  p-value: < 2.2e-16
```

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea7, 2020.
- [2] Gutiérrez B. Ana. *Probabilidad y Estadística, Enfoque por competencias*. McGraw Hill, 2012.
- [3] Mangiafico, Salvatore S. Summary and Analysis of Extension Program Evaluation in R. Recurso disponible en, https://rcompanion.org/handbook/I_12.html.
- [4] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 8

1. Discusión sobre material proporcionado en clase

El teorema de Bayes permite que el proveedor de atención médica convierta los resultados de una prueba en la probabilidad de tener una enfermedad. Este teorema se utiliza para calcular la probabilidad condicional de un suceso, teniendo información de antemano de dicho suceso, en otras palabras, sirve para determinar la probabilidad de una de las causas, puesto que ya se observó el efecto o suceso, y se calcula de la manera siguiente:

$$P(A_n | B) = \frac{P[B | A_n] * P[A_n]}{\sum_{i=1}^n P[B | A_i] * P[A_i]}, \quad (1)$$

donde B es el suceso sobre el que tenemos información previa, A_n son los distintos sucesos condicionados y A_i es la causa i , donde $i = 1, 2, 3, \dots, n$. Aplicando la ley de la multiplicación a la ecuación 1 se obtiene la siguiente expresión (ecuación 2):

$$P(A | B) = \frac{P[B | A] * P[A]}{P[B]}, \quad (2)$$

donde en la parte del numerador se tiene la probabilidad condicionada, y en la parte del denominador la probabilidad total.

De acuerdo a la enfermedad del Covid-19, hay diversos estudios donde se utiliza el teorema de Bayes para analizar los casos que se tienen al realizar la prueba y de los factores que influyen para su cálculo. Los casos que se pueden presentar son:

- Verdaderamente positivos: si una persona con Covid-19 da positivo a este.
- Verdaderamente negativos: cuando una persona sin Covid-19 da negativo en las pruebas.

-
- Falsos positivos: cuando una persona sin Covid–19 da positivo a este.
 - Falsos negativos: si una persona con Covid–19 da negativo en la prueba.

Para interpretar con una mejor precisión el resultado de la prueba, es necesario conocer su valor predictivo positivo (VPP)¹ y negativo (VPN)², los cuales dependen de la sensibilidad y especificidad y la prevalencia o la probabilidad previa a la prueba (probabilidad a priori). Se conoce como sensibilidad de cualquier prueba biomédica a la probabilidad de que una persona dé positivo dado que tiene la enfermedad y, la especificidad a la probabilidad de que una persona dé negativo dado que no tiene la enfermedad. En el trabajo de Schnipper [10], de manera interactiva muestran como calcular los VPP y VPN.

Por ejemplo, en la investigación de Lewis [8], se utiliza el teorema de Bayes para demostrar cómo la calidad de una prueba de Covid–19 depende de la magnitud del brote (tasa base), en este estudio, reescriben la ecuación 2 usando las probabilidades relevantes para las pruebas de Covid–19 dando como resultado la ecuación 3,

$$P(Cov | Pos) = \frac{P[Pos | Cov] * P[Cov]}{P[Pos | Cov] * P[Cov] + P[Pos | NoCov] * P[NoCov]}, \quad (3)$$

donde $P(Cov | Pos)$ es la probabilidad de que una persona tenga Covid–19 dado que ha dado positivo en la prueba, $P(Pos | Cov)$ es la sensibilidad de la prueba, $P(Cov)$ es la tasa base o la prevalencia (número de casos de Covid–19 / número total de la población a analizar), $P(NoCov)$ el porcentaje de personas que no tienen la enfermedad, $(1 - P(Cov))$ y $P(Pos | NoCov)$ es la probabilidad de los falsos positivos.

De acuerdo a la ecuación 3, considero que es correcto determinar que, incluso cuando se utiliza una prueba muy sensible, cuanto menor sea la tasa base de la enfermedad, más probabilidad se tiene de obtener falsos positivos. Asimismo, entre mayor sea la tasa base, más probabilidad hay de obtener falsos negativos a medida que se aumenten las pruebas.

Por otro lado, en el trabajo realizado por Ranjan [9], propone un modelo donde calculan el valor predictivo positivo VPP y la prevalencia o tasa base está determinada por la ecuación 4,

$$P = 1 - (1 - p)^x, \quad (4)$$

donde P es la probabilidad de que al menos 1 persona de las x cantidad de muestras tenga Covid–19 y p es la probabilidad de que 1 persona tenga la enfermedad. En este trabajo, el hecho de que la prevalencia sea igual al número de casos de Covid–19 dividido el número total de la población a analizar, demuestra que el cálculo del VPP, no es un indicador de gran ayuda para determinar el porcentaje de personas realmente están infectadas, aún sí se considera que las personas a evaluar viven en un área donde es más

¹El valor predictivo positivo (VPP) permite identificar el porcentaje de personas que tienen la enfermedad cuando la prueba ha dado positiva.

²El valor predictivo negativo (VPN) de una prueba es la “capacidad de descartar una enfermedad” dado un resultado negativo de la prueba.

probable que contraiga el virus (zona roja), por lo tanto, la metodología propuesta en cuanto a determinar los verdaderos positivos con base en el número de muestras realizadas, no parece erróneo. Sin embargo, considero que el estudio realizado carece de claridad respecto a los resultados obtenidos.

En la investigación de Good [6], aplican un análisis bayesiano (cálculo del VPP y VPN) para ilustrar la interpretación de las pruebas negativas de Covid-19 considerando que la prevalencia es la probabilidad previa a la prueba, es decir, una probabilidad con base en la sospecha clínica de padecer la enfermedad. Proponen dos escenarios (alta y baja probabilidad previa a la prueba de infección por Covid-19) clínicos. Para ambos escenarios, asumen una especificidad del 99,9 % y varían la sensibilidad del 70 % al 90 %. Considero que los resultados obtenidos van acorde a la metodología planteada, puesto que se realizó de manera correcta la interpretación de los VPN y VPP.

Finalmente, en la investigación de Chan [3], se realiza el estudio basado en las razones de verosimilitud negativa (LR_- , por sus siglas en inglés de, *negative likelihood ratio*) y de odds (razón de probabilidades), los cuales son posibles de calcular utilizando el teorema de Bayes y, la prevalencia o probabilidad previa de la prueba es una probabilidad a priori, es decir, con base en la experiencia del médico y la información que posee acerca de la enfermedad en particular del paciente. Este estudio fue realizado con el fin de obtener una mejor estimación del riesgo que tiene un determinado paciente en tener o contraer una enfermedad cuando el resultado de la prueba es negativo. Cabe aclarar que el cálculo de los LR dependen del tipo de análisis que se desee realizar [5]. El LR_- , los **odds** y la conversión de estos en **probabilidad**, se calculan mediante las ecuaciones 5, 6 y 7, respectivamente,

$$LR_- = \frac{1 - \text{sensibilidad}}{\text{especificidad}}, \quad (5)$$

$$\text{odds} = \frac{\text{probabilidad previa}}{1 - \text{probabilidad previa}}, \quad (6)$$

$$\text{probabilidad} = \frac{\text{odds}}{\text{odds} + 1}. \quad (7)$$

Sin embargo, en lugar de realizar los cálculos de las ecuaciones 6 y 7, se puede utilizar el nomograma de Fagan, el cual proporciona una estimación visual de las probabilidades posteriores a la prueba con base en los LR . En la figura 1 se muestra el nomograma de Fang basado en el estudio realizado por Aznar-Orovala [1].

En conclusión, un resultado negativo o positivo sin considerar los antecedentes clínicos, factores genéticos o datos que se conoce con precisión del paciente puede limitar la capacidad de los médicos para realizar acciones y disposiciones apropiadas. Para todas las pruebas de detección, ya sea para Covid-19 u otros diagnósticos, la comprensión de los valores predictivos y las razones de probabilidad con la ayuda del teorema de Bayes, podrían garantizar una interpretación sólida y las recomendaciones y acciones resultantes por parte de los médicos y las partes interesadas.

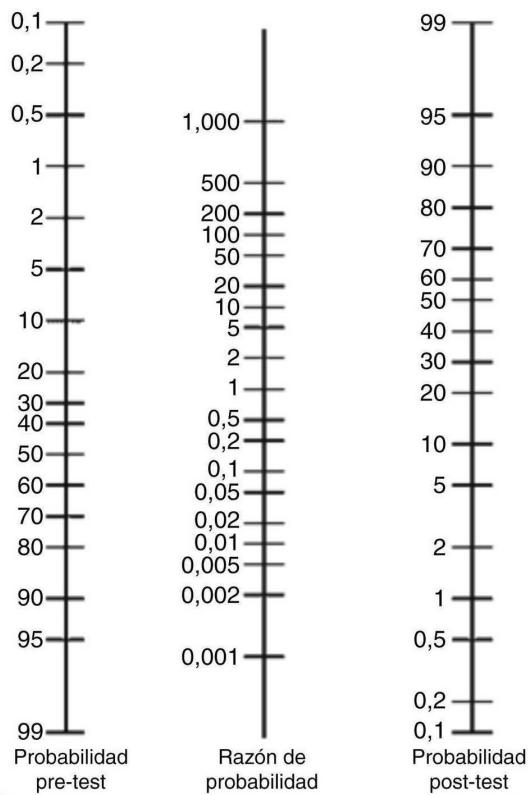


Figura 1: Nomograma de Fagan, donde línea izquierda son las probabilidades previas a la prueba, la linea central es el LR y línea derecha son las probabilidades de tener la enfermedad después del uso de la prueba.

2. Caso práctico

Para la aplicación del teorema de Bayes, en este trabajo se realizó un análisis para estimar de que tan probable es que un paciente con prueba positiva por Covid–19 pueda tener la enfermedad o si el resultado es negativo, que tan probable fue que no padeciera de ella, para lo cual se utilizan los valores predictivos positivos (VPP) o negativos (VPN). Para el cálculo de estos valores, se hace necesario conocer datos como la prevalencia estimada de la enfermedad, la sensibilidad y especificidad de la prueba. Dado que las pruebas por Covid–19 en México se han confirmaron mediante RT-PCR en tiempo real y, aproximadamente, a partir del mes de septiembre el IMMS empezó a utilizar pruebas rápidas en sus hospitales con el objetivo es diferenciar rápidamente el Covid–19 de la influenza, se investigó tanto la sensibilidad como especificidad que tienen ambas pruebas.

De acuerdo a Díaz-Jiménez [4], las pruebas PCR tienen una especificidad cercana al 100 % y una sensibilidad entre el 60 % y 80 % y, para las pruebas rápidas la sensibilidad más baja es del 20 %. Sin embargo, de acuerdo a las especificaciones de las pruebas PCR, se indica que la sensibilidad alcanza un 99 %. Para hallar la prevalencia, según la investigación de Lewis [8], se calcula de acuerdo a la magnitud de

brote y la determina mediante la ecuación 8. Por otro lado, en la investigación de Good [6], la prevalencia podría ser también una probabilidad previa a la prueba, es decir, una probabilidad de la enfermedad con base en la sospecha clínica,

$$\text{prevalencia} = \frac{\text{Total casos por Covid-19}}{\text{Población total}}. \quad (8)$$

Para efectos del presente trabajo, se consideraron dos escenarios con respecto al valor de la prevalencia. En el escenario 1, la prevalencia se calculó mediante la ecuación 8, por lo tanto, de acuerdo a la base de datos proporcionada por Hasell [7] a octubre 25 de 2020, hay 891,160 total de casos por Covid-19 y la población de México es de, aproximadamente, 128,932,753 habitantes, entonces, la prevalencia para el escenario 1 es igual a 0,69 % (0.0069).

Para el escenario 2, se contemplan dos prevalencias o probabilidades previas a la prueba por infección con Covid-19. Una probabilidad de prevalencia baja del 2 % y una alta del 80 %.

Para ambos escenarios se asumió una sensibilidad para las pruebas rápidas del 20 % y para las pruebas PCR la sensibilidad del 80 % y 99 %, La especificidad tanto para los escenarios, como para los tipos de pruebas varían entre el 99 % y 99,9 %. El cálculo los valores predictivos positivos y negativos se realizaron mediante las ecuaciones 9 y 10, respectivamente, basado en la investigación de [5],

$$VPP = \frac{\text{sensibilidad} * \text{prevalencia}}{\text{sensibilidad} * \text{prevalencia} + (1 - \text{sensibilidad}) * (1 - \text{especificidad})}, \quad (9)$$

$$VPN = \frac{\text{especificidad} * (1 - \text{prevalencia})}{(1 - \text{sensibilidad}) * \text{prevalencia} + \text{especificidad} * (1 - \text{prevalencia})}. \quad (10)$$

Es importante tener en cuenta que los VPP estiman la probabilidad condicional de que la persona pueda estar enferma cuando la prueba ha dado positiva y los VPN es la “capacidad de descartar una enfermedad” dado un resultado negativo en la prueba, por lo tanto, para determinar la probabilidad condicional de que una persona pueda estar enferma dado que su prueba es negativa (falso negativo) será igual a $1 - \text{VPN}$.

Considerando lo antes mencionado y los resultados del cuadro 1, observamos que al comparar la efectividad o sensibilidad del 99 % en un resultado positivo a través de las pruebas de PCR, con la probabilidad condicional, si se considera la prevalencia con respecto a la magnitud del brote (prevalencia muy baja (0.69 %), da como resultado una estimación de la efectividad de la prueba del 41 % y 87 %, con especificidades del 99 % y 99.9 %, respectivamente. Mientras que, si se considera una prevalencia alta, es decir, con base al riesgo al que ha estado expuesta la persona, la probabilidad condicional estimada para la efectividad de la prueba oscila entre el 96 % y 100 %. Por lo anterior, se podría llegar a un resultado contraintuitivo, ya que la probabilidad de que una persona con un resultado positivo a través de las pruebas PCR se pueda determinar como un posible caso verdadero con Covid-19, oscila entre un 41 % y 87 % y no con una probabilidad del 99 % como se ha especificado para este tipo de pruebas, aunque este resultado depende de la prevalencia considerada, como se observa en la figura 2.

Cuadro 1: Estimación de la probabilidad condicional para las pruebas por Covid-19

Escenario	Prevalencia (%)	Especificidad (%)	Sensibilidad (%)	Teorema de Bayes	
				Prueba positiva (%)	Prueba negativa (%)
1	0.69	99.0	20	12.2	0.6
			80	35.8	0.1
			99	40.8	0.0
		99.9	20	58.2	0.6
			80	84.8	0.1
			99	87.3	0.0
2	2.00	99.0	20	29.0	1.6
			80	62.0	0.4
			99	66.9	0.0
		99.9	20	80.3	1.6
			80	94.2	0.4
			99	95.3	0.0
	80.00	99.0	20	98.8	76.4
			80	99.7	44.7
			99	99.7	3.9
		99.9	20	99.9	76.2
			80	100.0	44.5
			99	100.0	3.8

De igual manera, cuando los resultados de las pruebas son negativos, se observa que mientras la prevalencia aumenta, también aumenta la probabilidad de falsos negativos, siendo, principalmente, un foco de atención las pruebas con sensibilidad muy baja, ya que se muestra una probabilidad muy alta (aproximadamente del 76 %) de que las personas con resultados negativos por Covid–19, no estén posiblemente sin la enfermedad, es decir, un aumento de la tasa en los falsos negativos. Por lo tanto, un resultado negativo obtenido por las pruebas rápidas (sensibilidad del 20 %), no sería un buen indicador para descartar la existencia de la enfermedad, ya que como podemos observar en el cuadro 1, con una prevalencia baja o alta, se presenta la probabilidad más alta de falsos negativos. No obstante, considerando que el resultado sea positivo, es posible que puedan ser de utilidad ya que la probabilidad condicional estimada es relativamente buena comparada las pruebas PCR a medida que aumenta la prevalencia, como se observa en la figura 2.

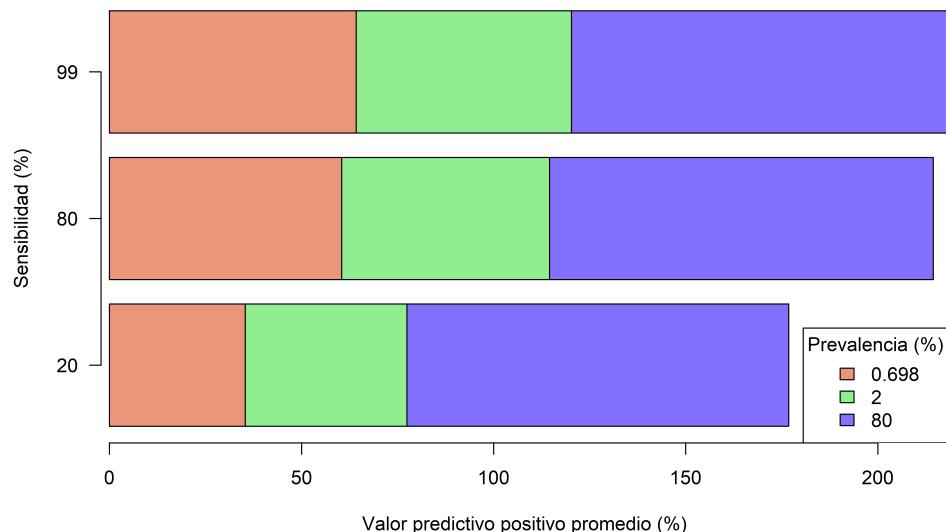


Figura 2: Valores predictivos positivos promedio

Como conclusión general, para las pruebas PCR, mediante el teorema de Bayes, es decir, por medio de los VPP y VPN si se considera un muestreo masivo de la población, determinar si la eficiencia de la prueba es alta, depende de la prevalencia real de la enfermedad del lugar donde se aplica, en este caso, de la magnitud del brote. Sin embargo, cuando el muestreo se considera con base a sectores donde se ha comprobado una mayor prevalencia o riesgo de contagio, la eficiencia (sensibilidad) reportada de la prueba ayuda a estimar con una mayor probabilidad la detección de casos en la población y, de esta manera aplicar las medidas de control necesarias. La base de datos y el código en R utilizado, se encuentran disponibles en el repositorio de GitHub [2].

Referencias

- [1] Aznar-Orovala, E., Mancheño-Alvarob, A., García-Lozanoa, T., Sánchez-Yepes, M. Likelihood ratio and Fagan's nomogram: 2 basic tools for the rational use of clinical laboratory tests. *Revista de Calidad Asistencial*, 28(6):390–391, 2013.
- [2] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea8, 2020.
- [3] Chan, Gar Ming. Interpreting COVID-19 Test Results: a Bayesian Approach. *Interpreting COVID-19 Test Results: a Bayesian Approach*, 06 2020.
- [4] Díaz-Jiménez, Irma Virginia. Interpretation of diagnostic tests for the SARS-Cov-2 virus. *Acta Pediatr Mex*, 41(Supl 1):S51–S57, 2020.
- [5] Fernández R., Raúl. Bayes's theorem and its use in diagnostic test lectures in clinical laboratory. *Revista Cubana de Investigaciones Biomédicas*, 28(3):158–165, 2009.
- [6] Good, Chester B., Hernandez, Inmaculada, Smith, Kenneth. Interpreting COVID-19 Test Results: a Bayesian Approach. *Journal of General Internal Medicine*, 35(8):2490–2491, 06 2020.
- [7] Hasell, J., Mathieu, E., Beltekian, D. A cross-country database of COVID-19 testing. *Sci Data* 7, 345. Recurso disponible en, <https://ourworldindata.org/coronavirus-testing>, 2020.
- [8] Lewis, Michael A. Bayes' theorem and Covid-19 testing. *Significance*, 04 2020.
- [9] Ranjan, Archit. COVID-19, Bayes' theorem and taking probabilistic decisions. Recurso disponible en, <https://towardsdatascience.com/covid-19-bayes-theorem-and-taking-data-driven-decisions-part-1-b61e2c2b3bea>, 2020.
- [10] Schnipper, Effrey L., Sax, Paul E. Covid-19 test accuracy supplement: The math of Bayes' Theorem. Recurso disponible en, <https://www.statnews.com/2020/08/20/covid-19-test-accuracy-supplement-the-math-of-bayes-theorem/>, 2020.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 9

1. Problemas a resolver

En el presente trabajo se realizaron las soluciones de diversos problemas del libro de Grinstead [1] sobre el valor esperado, varianza y desviación estándar de una variable aleatoria (v.a) discreta y continua. Para el cálculo de estos valores se utilizaron las ecuaciones 1, 2, 3 y 4, donde X es una v.a con espacio muestral Ω y distribución de probabilidad $f(x)$:

Valor esperado si X es discreta

$$\mu = E(X) = \sum_{x \in \Omega} xf(x). \quad (1)$$

Valor esperado si X es continua,

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx. \quad (2)$$

Varianza de X ,

$$\sigma^2 = V(X) = E(X^2) - \mu^2 \quad (3)$$

Desviación estándar de X ,

$$\sigma = D(X) = \sqrt{V(X)}. \quad (4)$$

1.1. Problema 1, página 247

A card is drawn at random from a deck consisting of cards numbered 2 through 10. A player wins 1 dollar if the number on the card is odd and loses 1 dollar if the number is even. What is the expected value of his winnings?

Solución

Numeración de las cartas: 2, 3, 4, 5, 6, 7, 8, 9, 10.

Cantidad total de cartas = 9, donde hay 5 cartas pares y 4 impares. Si saca una carta par, gana 1 dólar, de lo contrario, pierde 1 dólar. Entonces, sea X el evento de que la carta sacada sea par o impar, el valor esperado de ganancia se calcula mediante la ecuación 1:

$$\begin{aligned} E(X) &= 1 \left(\frac{5}{9} \right) - 1 \left(\frac{4}{9} \right) \\ E(X) &= \frac{1}{9}. \end{aligned}$$

1.2. Problema 6, página 247

A die is rolled twice. Let X denote the sum of the two numbers that turn up, and Y the difference of the numbers (specifically, the number on the first roll minus the number on the second). Show that $E(XY) = E(X)E(Y)$. Are X and Y independent?

Solución

$X = a + b$, $Y = a - b$, donde a es el primer lanzamiento y b el segundo. En total hay 36 combinaciones posibles al lanzar dos veces el dado.

Utilizando la ecuación 1, encuentro el valor esperado de la variable aleatoria Y :

$$\begin{aligned} E(Y) &= -1 \left(\frac{5}{36} \right) - 2 \left(\frac{4}{36} \right) - 3 \left(\frac{3}{36} \right) - 4 \left(\frac{2}{36} \right) - 5 \left(\frac{1}{36} \right) + 1 \left(\frac{5}{36} \right) \\ &\quad + 2 \left(\frac{4}{36} \right) - 3 \left(\frac{3}{36} \right) + 4 \left(\frac{2}{36} \right) + 5 \left(\frac{1}{36} \right) + 0 \left(\frac{6}{36} \right) \\ E(Y) &= \frac{-1 - 2 - 3 - 4 - 5 + 1 + 2 + 3 + 4 + 5}{36} \\ E(Y) &= E(a - b) = 0. \end{aligned}$$

Entonces, $E(X)E(Y) = E(X) * 0 = 0$

Ahora, ¿ $E(XY) = 0$?

$$\begin{aligned} E(XY) &= E(a + b) * E(a - b) \\ E(XY) &= E(a^2 - ab + ba - b^2) \\ E(XY) &= E(a^2 - b^2) = 0. \end{aligned}$$

Por lo tanto, $E(XY) = E(X)E(Y) = 0$.

¿Son X y Y variables aleatorias independientes? Para comprobar que lo son o no, basta con encontrar un caso en que no se cumpla la independencia, es decir, $P(X, Y) = P(X)P(Y)$, por ejemplo,

si $X = 12$ y $Y = 0$ y la probabilidad de sacar cualquier cara del dado es de 1/6:

$$P(X = 12, Y = 0) = P(a = 6, b = 6)$$
$$P(X = 12, Y = 0) = \frac{1}{36}.$$

Ahora, ¿ $P(X)P(Y) = 1/36$?

$$P(X = 12) = \frac{1}{36}$$
$$P(Y = 0) = \frac{6}{36}$$
$$P(X = 12) * P(Y = 0) = \frac{1}{36} * \frac{1}{6}$$
$$P(X = 12) * P(Y = 0) = \frac{1}{216}.$$

Entonces, $P(X = 12, Y = 0) \neq P(X = 12) * P(Y = 0)$, por lo tanto, X y Y no son variables aleatorias independientes.

1.3. Problema 15, página 249

A box contains two gold balls and three silver balls. You are allowed to choose successively balls from the box at random. You win 1 dollar each time you draw a gold ball and lose 1 dollar each time you draw a silver ball. After a draw, the ball is not replaced. Show that, if you draw until you are ahead by 1 dollar or until there are no more gold balls, this is a favorable game.

Solución

Sea D las bolas doradas y P las bolas plateadas, en la caja hay un total de 5 bolas, $2D$ y $5P$. Entonces, las formas de ganar sería si sacara las bolas en el siguiente orden:

- La primera bola en sacar es D , entonces, su probabilidad sería $2/5$.
- PDD , entonces, su probabilidad sería $3/5 * 2/4 * 1/3 = 1/10$.

Las formas de perder 1 dólar serían:

- $PPPDD$, entonces, su probabilidad sería $3/5 * 2/4 * 1/3 * 1/2 * 1 = 1/10$.
- $PDPPD$, entonces, su probabilidad sería $3/5 * 2/4 * 2/3 * 1/2 * 1 = 1/10$.
- $PPDPD$, entonces, su probabilidad sería $3/5 * 2/4 * 2/3 * 1/2 * 1 = 1/10$.

Por lo tanto, sea X el evento de ganar o perder en el juego, utilizando la ecuación 1, se encuentra el valor esperado de la variable aleatoria X , considerando que el valor por ganar es de 1 dolar y de -1 si pierde,

no se considera cuando empata porqué su valor sería de 0, entonces se tiene:

$$E(X) = 1 \left(\frac{2}{5} \right) + 1 \left(\frac{1}{10} \right) - 1 \left(\frac{1}{10} \right) - 1 \left(\frac{1}{10} \right) - 1 \left(\frac{1}{10} \right)$$
$$E(X) = \frac{2}{10} = \frac{1}{5}.$$

1.4. Problema 18, página 249

Exactly one of six similar keys opens a certain door. If you try the keys, one after another, what is the expected number of keys that you will have to try before success?

Solución

Sea X el número de intentos antes de encontrar la llave correcta, los intentos F antes de encontrar la llave correcta C serían los siguientes:

- 0 intentos: C , entonces, su probabilidad sería $1/6$.
- 1 intento: FC , entonces, su probabilidad sería $5/6 * 1/5 = 1/6$.
- 2 intentos: FFC , entonces, su probabilidad sería $5/6 * 4/5 * 1/4 = 1/6$.
- 3 intentos: $FFFC$, entonces, su probabilidad sería $5/6 * 4/5 * 3/4 * 1/3 = 1/6$.
- 4 intentos: $FFFFC$, entonces, su probabilidad sería $5/6 * 4/5 * 3/4 * 2/3 * 1/2 = 1/6$.
- 5 intentos: $FFFFC$, entonces, su probabilidad sería $5/6 * 4/5 * 3/4 * 2/3 * 1/2 = 1/6$.

Por lo tanto, el valor esperado de la v.a X la encontramos por medio de la ecuación 1:

$$E(X) = 0 \left(\frac{1}{6} \right) + 1 \left(\frac{1}{6} \right) + 2 \left(\frac{1}{6} \right) + 3 \left(\frac{1}{6} \right) + 4 \left(\frac{1}{6} \right) + 5 \left(\frac{1}{6} \right)$$
$$E(X) = \frac{15}{6} = \frac{5}{2}.$$

1.5. Problema 19, página 249

A multiple choice exam is given. A problem has four possible answers, and exactly one answer is correct. The student is allowed to choose a subset of the four possible answers as his answer. If his chosen subset contains the correct answer, the student receives three points, but he loses one point for each wrong answer in his chosen subset. Show that if he just guesses a subset uniformly and randomly his expected score is zero.

Solución

De las 4 posibles respuestas (A, B, C, D), el alumno podría escoger, 0, 1, 2, 3 o 4 subconjuntos de preguntas como su respuesta, y en caso de estar la respuesta correcta dentro de cada subconjunto escogido, el estudiante ganará 3 puntos, pero perderá 1 por cada pregunta incorrecta, sea X las posibles respuestas que selecciona el alumno, el valor esperado para cada subconjunto se calcula con la ecuación 1:

- Subconjunto con 0 respuestas, no perdería ni ganaría puntos, por lo tanto, $E(X) = 0$.
- Subconjunto con 1 respuesta, A, B, C o D:

$$E(X) = 3(1) - 3(1)$$

$$E(X) = 0.$$

- Subconjunto con 2 respuestas, A-B, A-C, A-D, B-C, B-D o C-D:

$$E(X) = (3 - 1) \left(\frac{3}{6}\right) - 2 \left(\frac{3}{6}\right)$$

$$E(X) = 0.$$

- Subconjunto con 3 posibles respuestas, A-B-C, A-B-D, A-C-D, B-D-C:

$$E(X) = (3 - 2) \left(\frac{3}{4}\right) - 3 \left(\frac{1}{4}\right)$$

$$E(X) = 0.$$

- Subconjunto con 4 posibles respuestas, A-B-C-D:

$$E(X) = 3 \left(\frac{1}{4}\right) - 1 \left(\frac{3}{4}\right)$$

$$E(X) = 0.$$

En todos los Subconjunto en valor esperado es 0.

1.6. Problema 1, página 263

A number is chosen at random from the set $S = \{-1, 0, 1\}$. Let X be the number chosen. Find the expected value, variance, and standard deviation of X .

Solución

Sea X el número escogido, por lo tanto, $p(x) = 1/3$, el valor esperado de X se calcula con la ecuación 1, su varianza con la ecuación 3 y desviación estándar con la ecuación 4:

$$\mu = E(X) = -1 \left(\frac{1}{3}\right) + 0 \left(\frac{1}{3}\right) + 1 \left(\frac{1}{3}\right)$$

$$\mu = E(X) = 0.$$

Ahora, para calcular la varianza, se necesita el valor de $E(X^2)$, entonces:

$$E(X^2) = (-1)^2 \left(\frac{1}{3}\right) + (0)^2 \left(\frac{1}{3}\right) + (1)^2 \left(\frac{1}{3}\right)$$

$$E(X^2) = \frac{2}{3}.$$

Por lo tanto, la varianza y desviación estándar serán:

$$V(X) = \left(\frac{2}{3}\right) - (0)^2 \quad (5)$$

$$V(X) = \frac{2}{3}.$$

$$D(X) = \sqrt{V(X)}$$

$$D(X) \sqrt{\frac{2}{3}} \approx 0.816.$$

1.7. Problema 9, página 264

A die is loaded so that the probability of a face coming up is proportional to the number on that face. The die is rolled with outcome X . Find $V(X)$ and $D(X)$.

Solución

Sea k la proporción de la cara del dado cargado, entonces, la probabilidad de cada cara sería:

$P(1) = \frac{1}{k}$, $P(2) = \frac{2}{k}$, $P(3) = \frac{3}{k}$, $P(4) = \frac{4}{k}$, $P(5) = \frac{5}{k}$, $P(6) = \frac{6}{k}$ y la suma de estas proporciones debe ser 1, entonces:

$$1 = \frac{1}{k} + \frac{2}{k} + \frac{3}{k} + \frac{4}{k} + \frac{5}{k} + \frac{6}{k}$$

$$k = 21.$$

El valor esperado de X se calcula con la ecuación 1, su varianza con la ecuación 3 y desviación estándar con la ecuación 4:

$$\mu = E(X) = 1 \left(\frac{1}{21}\right) + 2 \left(\frac{2}{21}\right) + 3 \left(\frac{3}{21}\right) + 4 \left(\frac{4}{21}\right) + 5 \left(\frac{5}{21}\right) + 6 \left(\frac{6}{21}\right)$$

$$\mu = E(X) = \frac{1 + 4 + 9 + 16 + 25 + 36}{21}$$

$$\mu = E(X) = \frac{13}{3}.$$

Ahora, para calcular la varianza, se necesita el valor de $E(X^2)$, entonces:

$$\begin{aligned} E(X^2) &= (1)^2 \left(\frac{1}{21} \right) + (2)^2 \left(\frac{2}{21} \right) + (3)^2 \left(\frac{3}{21} \right) + (4)^2 \left(\frac{4}{21} \right) + (5)^2 \left(\frac{5}{21} \right) + (6)^2 \left(\frac{6}{21} \right) \\ E(X^2) &= 1 \left(\frac{1}{21} \right) + 4 \left(\frac{2}{21} \right) + 9 \left(\frac{3}{21} \right) + 16 \left(\frac{4}{21} \right) + 25 \left(\frac{5}{21} \right) + 36 \left(\frac{6}{21} \right) \\ E(X^2) &= \frac{1 + 8 + 27 + 64 + 125 + 216}{21} \\ E(X^2) &= 21 \end{aligned}$$

Por lo tanto, la varianza y desviación estándar serán:

$$\begin{aligned} V(X) &= 21 - \left(\frac{13}{3} \right)^2 \\ V(X) &= \frac{20}{9}. \\ D(X) &= \sqrt{V(X)} \\ D(X) &= \sqrt{\frac{20}{9}} \approx 1.49. \end{aligned}$$

1.8. Problema 12, página 264

Let X be a random variable with $\mu = E(X)$ and $\sigma^2 = V(X)$. Define $X^* = (X - \mu)/\sigma$. The random variable X^* is called the standardized random variable associated with X . Show that this standardized random variable has expected value 0 and variance 1.

Solución

$$\begin{aligned} E(X^*) &= E \left(\frac{X - \mu}{\sigma} \right) \\ &= \frac{1}{\sigma} E(X - \mu) \\ &= \frac{1}{\sigma} [E(X) - E(\mu)] \\ &= \frac{1}{\sigma} [\mu - \mu] \\ &= 0 \end{aligned}$$

La varianza se calcula por medio de la ecuación 3:

$$\begin{aligned} V(X^*) &= E(X^{*2}) - E(X^*)^2 \\ &= E \left[\left(\frac{X - \mu}{\sigma} \right)^2 \right] - 0 \\ &= \frac{1}{\sigma^2} E[(X - \mu)^2] \\ &= \frac{\sigma^2}{\sigma^2} \\ &= 1 \end{aligned}$$

1.9. Problema 3, página 278

The lifetime, measure in hours, of the ACME super light bulb is a random variable T with density function $f_T(t) = \lambda^2 t e^{-\lambda t}$, where $\lambda = 0.05$. What is the expected lifetime of this light bulb? What is its variance?

Solución

El valor esperado se calcula mediante la ecuación 2, entonces:

$$\begin{aligned}\mu &= E(T) = \int_0^\infty t f_T(t) dt \\&= \int_0^\infty t(\lambda^2 t e^{-\lambda t}) dt \\&= \int_0^\infty t^2 \lambda^2 e^{-\lambda t} dt \longrightarrow U = \lambda t, dt = \frac{dU}{\lambda} \\&= \int_0^\infty \frac{U^2 e^{-U}}{\lambda} dU \\&= \frac{1}{\lambda} \int_0^\infty U^2 e^{-U} dU \implies AB - \int_0^\infty B dA, \quad A = U^2, B = -e^{-u} \\&= \frac{1}{\lambda} \left[(AB)|_0^\infty - \int_0^\infty B dA \right] \longrightarrow dA = 2U dU \\&= \frac{1}{\lambda} \left[(AB)|_0^\infty - \int_0^\infty -e^{-u} 2U dU \right] \\&= \frac{1}{\lambda} \left[(AB)|_0^\infty + 2 \left(\int_0^\infty e^{-U} U dU \right) \right] \implies PQ - \int_0^\infty Q dp \longrightarrow P = U, Q = e^{-U}, dP = dU \\&= \frac{1}{\lambda} \left[(AB)|_0^\infty + 2 \left(PQ - \int_0^\infty Q dU \right) \right] \\&= \frac{1}{\lambda} \left[(AB)|_0^\infty + 2 \left(PQ - e^{-U} \right)|_0^\infty \right] \longrightarrow P = U, Q = -e^{-U} \\&= \frac{1}{\lambda} \left[(AB)|_0^\infty + 2 \left(U e^{-U} - e^{-U} \right)|_0^\infty \right] \longrightarrow A = U^2, B = -e^{-u} \\&= \frac{1}{\lambda} \left[(-U^2 e^{-U})|_0^\infty + (-2U e^{-U} - 2e^{-U})|_0^\infty \right] \longrightarrow U = \lambda t \\&= \frac{1}{\lambda} \left[(-e^{-U}(U^2 + 2U + 2))|_0^\infty \right] \\&= \frac{2}{\lambda} = \frac{2}{0.05} \\&= 40.\end{aligned}$$

Ahora, para calcular la varianza, se necesita el valor de $E(T^2)$, entonces:

$$\begin{aligned}
E(T^2) &= \int_0^\infty t^2 f_T(t) dt \\
&= \int_0^\infty t^2 (\lambda^2 t e^{-\lambda t}) dt \\
&= \int_0^\infty t^3 \lambda^2 e^{-\lambda t} dt \\
&= \frac{1}{\lambda} \int_0^\infty t^3 \lambda^3 e^{-\lambda t} dt \longrightarrow U = \lambda t, dt = \frac{dU}{\lambda} \\
&= \frac{1}{\lambda^2} \int_0^\infty U^3 e^{-U} dU \implies AB - \int_0^\infty B dA, \quad A = U^3, B = -e^{-u} \\
&= \frac{1}{\lambda} \left[(AB)|_0^\infty - \int_0^\infty B dA \right] \longrightarrow dA = 3U dU \\
&= \frac{1}{\lambda} \left[(AB)|_0^\infty - \int_0^\infty B 3U dU \right] \longrightarrow A = U^3, B = -e^{-u} \\
&= \frac{1}{\lambda^2} \left[(-U^3 e^{-U})|_0^\infty - \int_0^\infty -e^{-u} 3U^2 dU \right] \\
&= \frac{1}{\lambda^2} \left[(-U^3 e^{-U})|_0^\infty + 3 \left(\int_0^\infty \textcolor{red}{U^2 e^{-u} dU} \right) \right] \\
&= \frac{1}{\lambda^2} [(-U^3 e^{-U})|_0^\infty + 3(2)] \\
&= \frac{6}{\lambda^2} = \frac{6}{(0.05)^2} \\
&= 2,400
\end{aligned}$$

Por lo tanto, para calcular la varianza se utiliza la ecuación 3:

$$\begin{aligned}
V(T) &= 2,400 - (40)^2 \\
&= 2,400 - 1,600 \\
&= 800
\end{aligned} \tag{6}$$

Referencias

- [1] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability*. American Mathematical Society, 2006.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 10

1. Simulación de problemas

En el presente trabajo se realizaron algunas simulaciones para comparar los resultados analíticos de las soluciones de diversos problemas del libro de Grinstead [2] sobre el valor esperado, varianza y desviación estándar de una variable aleatoria (v.a) discreta y continua. Para el cálculo de estos valores se utilizaron las ecuaciones 1, 2, 3 y 4, donde X es una v.a con espacio muestral Ω y distribución de probabilidad $f(x)$:

Valor esperado si X es discreta

$$\mu = E(X) = \sum_{x \in \Omega} xf(x). \quad (1)$$

Valor esperado si X es continua,

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx. \quad (2)$$

Varianza de X ,

$$\sigma^2 = V(X) = E(X^2) - \mu^2. \quad (3)$$

Desviación estándar de X ,

$$\sigma = D(X) = \sqrt{V(X)}. \quad (4)$$

Todas las simulaciones fueron realizadas en el software R versión 4.0.2 [3] y el código empleado se encuentra en el repositorio GitHub [1].

1.1. Problema 1, página 247

A card is drawn at random from a deck consisting of cards numbered 2 through 10. A player wins 1 dollar if the number on the card is odd and loses 1 dollar if the number is even. What is the expected value of his winnings?

Solución

Cantidad total de cartas = 9, donde hay 5 cartas pares y 4 impares. Si se saca una carta par, se pierde 1 dólar, de lo contrario, se gana 1 dólar. Entonces, sea X el evento de que la carta sacada sea par o impar. El valor esperado de ganancia fue calculado mediante la ecuación 1 dando como resultado un $E(X) = -\frac{1}{9} \approx -0.11$.

Con el fin de identificar si este juego es desfavorable con un $E(x) \approx -0.11$, se utilizó la función `sample()` para realizar una simulación donde se saca 1 carta en la primera jugada, luego 2 cartas en la segunda jugada y así sucesivamente hasta llegar a 100 jugadas (es decir, se sacan 100 cartas), se calcula el valor esperado en cada jugada y se contabilizó la frecuencia de estos valores.

En la figura 1a, se muestran los resultados obtenidos de esta simulación, en la que podemos observar que, de las 100 jugadas, la mayoría de veces (79 jugadas) este no es un juego favorable y en promedio el valor esperados más frecuentes es -0.1 . De igual forma, se realizó la simulación desde 1 hasta 1,000 jugadas (ver figura 1b) y el resultado fue el mismo, en la mayoría de veces (960 jugadas) arrojó un valor esperado negativo, siendo en promedio el más frecuentes de -0.1 . Por lo anterior, se concluye que tanto analíticamente como experimentalmente, el juego de las cartas descrito en el problema 1.1, es un juego desfavorable.

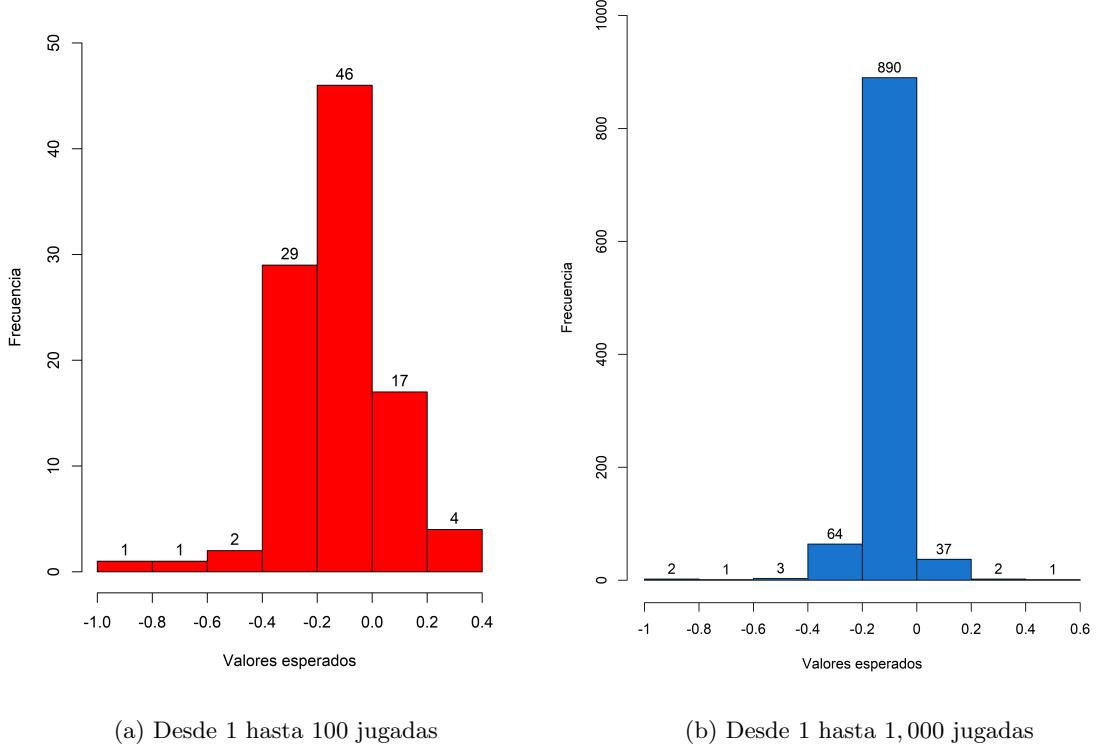
1.2. Problema 9, página 264

A die is loaded so that the probability of a face coming up is proportional to the number on that face. The die is rolled with outcome X . Find $V(X)$ and $D(X)$.

Solución

Se tiene que la probabilidad de cada cara de este dado cargado es de $p(x) = \frac{x}{21}$. El valor esperado, la varianza y desviación estándar fueron calculados mediante las ecuaciones 1, 3 y 4, respectivamente, las cuales arrojaron como resultado que el $E(X) = \frac{13}{3} \approx 4.33$, la $V(X) = \frac{20}{9} \approx 2.22$ y una $D(X) \approx 1.49$.

Se realizó una simulación para comparar si el valor esperado, varianza y desviación estándar calculados anteriormente, es el mismo para n lanzamientos. Se utilizó la función `sample()` para generar conjuntos con 10, 11, 12 y así sucesivamente, hasta de 100 lanzamientos, se calculó el valor esperado,



(a) Desde 1 hasta 100 jugadas

(b) Desde 1 hasta 1,000 jugadas

Figura 1: Resultados simulación del valor esperado para el problema 1.1

varianza y desviación estándar correspondiente en cada conjunto y se contabilizó la frecuencia de estos valores. Los resultados de esta simulación se muestran en la figura 2.

En la figura 2a, podemos observar que de los 91 conjuntos de lanzamientos el valor esperado más frecuente es, aproximadamente, de 4.3, en la figura 2b, el valor de la varianza más frecuente está alrededor de 2.2 y en la figura 2c el valor de la desviación estándar más frecuente en promedio es de 1.45. Por lo anterior, se puede concluir que tanto analíticamente como experimentalmente, en este dado cargado las caras con mayor probabilidad de salir son la 4, 5 y 6 con una desviación estándar de, aproximadamente, 1.45 como se puede observar en la figura 3.

1.3. Problema 1, página 263

A number is chosen at random from the set $S = \{-1, 0, 1\}$. Let X be the number chosen. Find the expected value, variance, and standard deviation of X .

Solución

Sea X el número escogido, por lo tanto, $p(x) = 1/3$, el valor esperado de X , la varianza y desviación

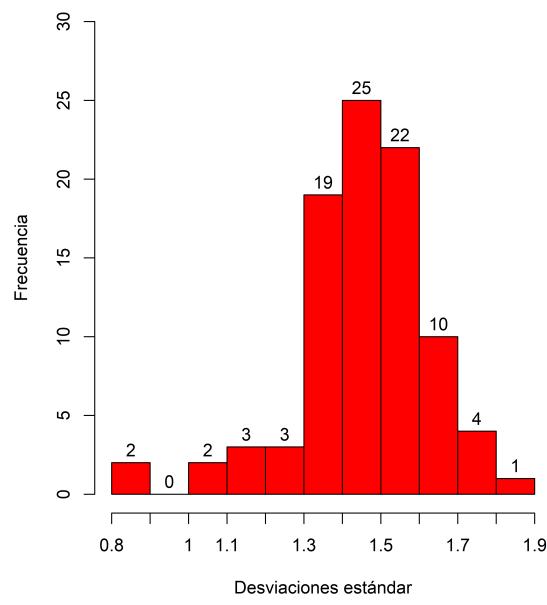
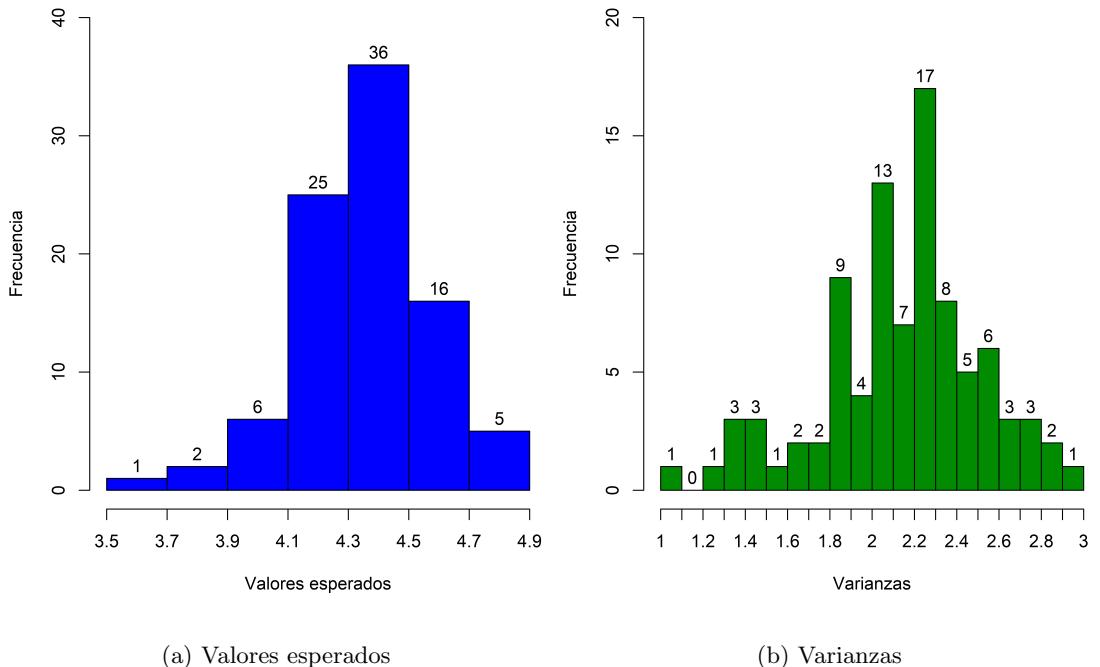


Figura 2: Resultados de la simulación del problema 1.2

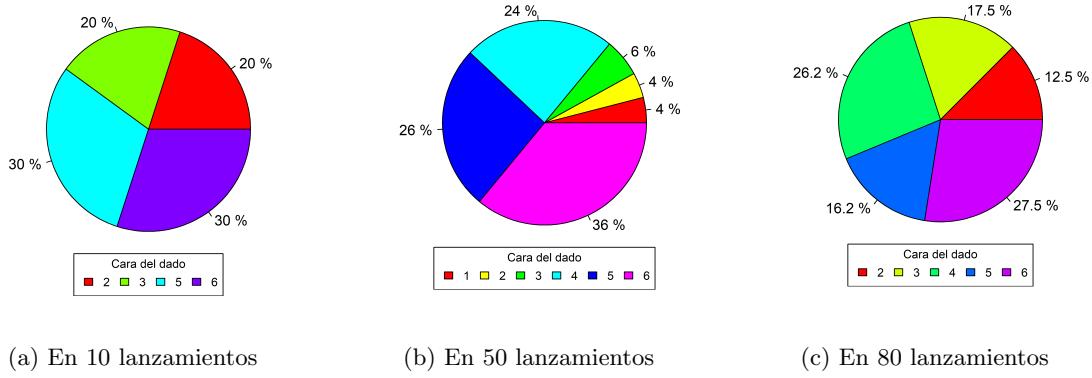


Figura 3: Resultado de la frecuencia de ocurrencias de las caras del dado cargado del problema 1.2

estándar fueron calculados mediante las ecuaciones 1, 3 y 4, las cuales dieron como resultado un $E(X) = 0$, una $V(X) = \frac{2}{3} \approx 0.66$ y una $D(X) \approx 0.816$.

Para este problema, en la simulación se consideró realizar 10,000 selecciones entre los números -1 , 0 y 1 y se calculó el valor esperado, varianza y desviación estándar con los datos obtenidos. Como resultado de esta simulación se obtuvo un valor esperado de -0.002 , una varianza de 0.670 y una desviación estándar de 0.818 , lo cual es muy similar a los valores obtenidos de manera analítica. En la figura 4 se muestra la frecuencia de selección de X .

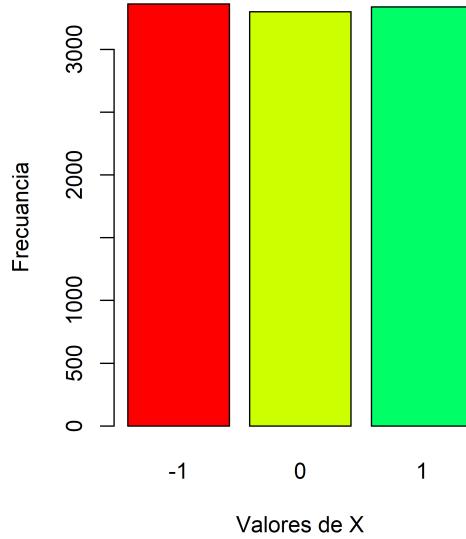


Figura 4: Frecuencia de selección entre los números -1 , 0 y 1 (problema 1.3)

1.4. Problema 3, página 278

The lifetime, measure in hours, of the ACME super light bulb is a random variable T with density function $f_T(t) = \lambda^2 t e^{-\lambda t}$, where $\lambda = 0.05$. What is the expected lifetime of this light bulb? What is its variance?

Solución

El valor esperado y varianza de la vida útil de las bombillas ACME se calcularon mediante la ecuación 2 y 3, las cuales arrojaron un $E(X) = 40$ y una $V(X) = 2.400$. Se utilizó la función `integrate()` para realizar estos cálculos.

Para este ejercicio, se realizó una simulación para determinar si el valor esperado y varianza de la vida útil de un lote de 100 bombillas que siguen una determinada distribución y con función de densidad $f_T(t)$ descrita en el problema 1.4, son iguales o similares a los valores calculados anteriormente. Los resultados de esta experimentación son mostrados en la figura 5.

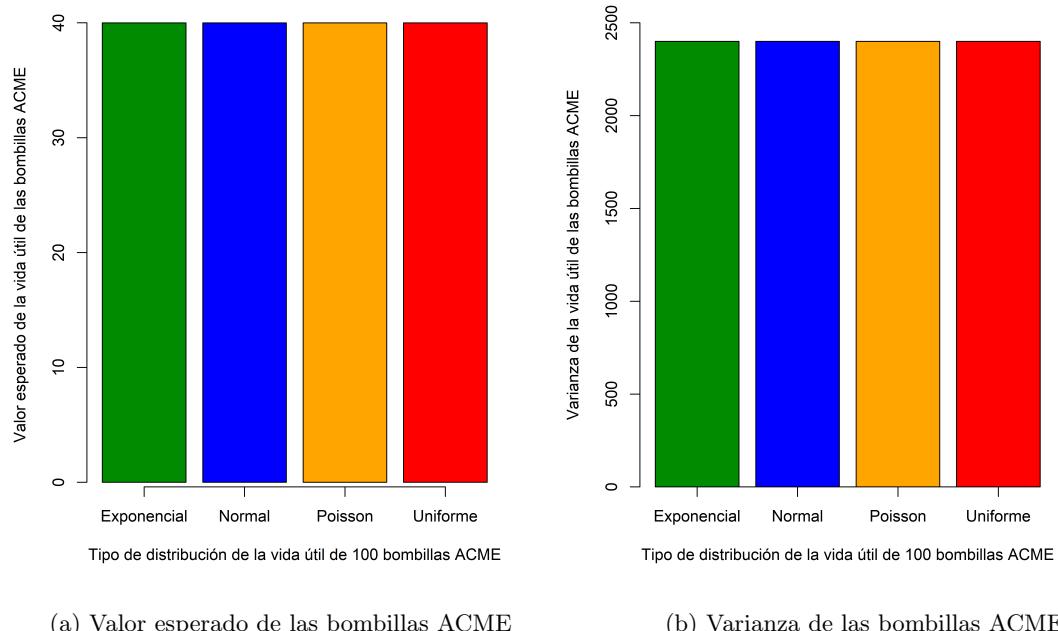


Figura 5: Resultados simulación para el problema 1.4

De acuerdo a lo anterior, podemos observar que en la figura 5a y 5b, sin importar el tipo de distribución que sigue la vida útil del lote de 100 bombillas ACME (en este caso, una distribución exponencial, uniforme, normal y de Poisson con valores positivos) el valor esperado y varianza, respectivamente, de la vida útil de estas bombillas es igual al calculado analíticamente, es decir $E(X) = 40$ y su $V(X) = 2.400$.

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea10, 2020.
- [2] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability*. American Mathematical Society, 2006.
- [3] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 11

1. Aplicación de la convolución

La convolución es un operador matemático que transforma dos funciones en una tercera función que representa la magnitud en la que se superponen una función sobre una versión trasladada e invertida de la otra función. De acuerdo a Kim [3], la convolución se puede describir como una función que es la integral (cuando las variables son continuas) o la suma de funciones (cuando las variables son discretas) de dos variables aleatorias independientes, y mide la cantidad de superposición cuando una función se desplaza sobre la otra. En las definiciones matemáticas la convolución, tanto discreta como continua, está indicada por el operador $*$.

En el procesamiento de señales digitales (DSP por sus siglas en inglés de *digital signal processing*), donde una de sus aplicaciones prácticas son el tratamiento de las imágenes, en la que la convolución juega un papel importante, ya que es una técnica de procesamiento de imágenes común que cambia las intensidades de un píxel para reflejar las intensidades de los píxeles circundantes. Además, se puede obtener efectos de imagen populares como desenfoque, nitidez y detección de bordes [4]. Un detector de bordes puede procesar y extraer características relevantes en un conjunto de imágenes antes de que se introduzcan en un algoritmo de reconocimiento de patrones, lo que puede dar como resultado un rendimiento superior, por ejemplo, que un automóvil sin conductor frena en una señal de alto. La mejora de imágenes puede resultar especialmente útil cuando se trata de imágenes científicas.

En el procesamiento de imágenes, muchas operaciones de filtro se aplican a una imagen realizando una operación especial llamada convolución con una matriz que recibe el nombre de kernel, los cuales son típicamente matrices cuadradas que van desde 2×2 , hasta de 5×5 , siendo la comúnmente utilizada la de 3×3 . Los valores almacenados en el kernel se relacionan directamente con los resultados de la aplicación

del filtro a la imagen y son los que determinan cómo transformar los píxeles de la imagen original en los píxeles de la imagen procesada [4].

Dado que las imágenes también se pueden considerar como cuadrículas bidimensionales de números, la aplicación de un kernel a una imagen se puede visualizar como una pequeña cuadrícula (el kernel) que se mueve a través de una cuadrícula sustancialmente más grande (la imagen), donde la convolución del kernel hace que el valor de cada píxel se recalcula utilizando la suma de vecindad ponderada definida en la matriz del kernel. La representación de los pasos de esta convolución se puede observar en la figura 1.

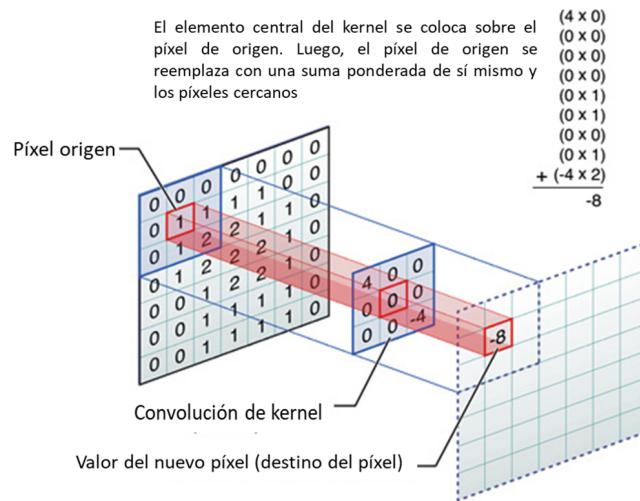


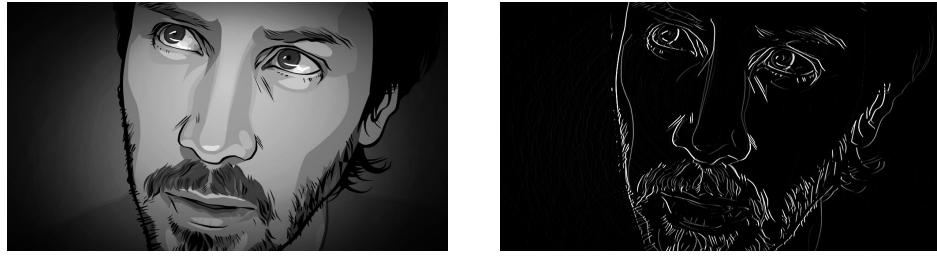
Figura 1: Ejemplo de convolución del kernel, tomado de la información de la referencia [4]

En la investigación desarrollada por Kim [3], realizan el tratamiento de imágenes a blanco y negro (escala de grises) mediante la convolución del kernel *Sobel*, el cual permite detectar los bordes verticales y horizontales en una imagen, obteniendo así el relieve de la imagen. En el programa R, mediante la función `magik()`, se pueden realizar el procesamiento a las imágenes y con la función `image_convolve()` se pueden realizar la convoluciones de kernel. En la figura 2 se muestra el resultado de aplicar la convolución con el kernel *Sobel* a una imagen. En la documentación de Thyssen [7] se encuentran más ejemplos de `convolve` con varios kernels disponibles. Además de aplicaciones como el relieve de la imagen, se pueden utilizar filtros para eliminar señales e imágenes (ruido). Diferentes filtros pueden lograr este propósito y el óptimo a menudo, depende de los requisitos particulares de la aplicación que se requiera.

De igual manera, en el trabajo de Kim [3], presentan las expresiones matemáticas para la convolución con imágenes, determinando que, para el caso de las señales digitales, si A_n y B_n son funciones con respecto a un valor n entero, entonces la convolución en un sistema unidimensional, se representa mediante la ecuación 1:

$$A[n_a] * B[n_b] = C[n_c] = \sum_{\tau=0}^{n_a-1} A[\tau] \times B[n_c - \tau], \quad (1)$$

donde $0 \leq n_c < n_a + n_b - 1$ y τ es la variable sobre la cual se realiza el desplazamiento.



(a) Imagen antes de la convolución.
 (b) Imagen después de la convolución con el
 kernel *Sobel*
 Fuente: *A Scanner Darkly (2006)*.

Figura 2: Ejemplo de imagen aplicando convolución de kernel

Sin embargo, debido a que las imágenes son bidimensionales, es necesario la extensión a 2-D de la ecuación 1 para realizar la convolución con imágenes, la cuál es mostrada en la ecuación 2:

$$A[i_a, j_a] * B[i_b, j_b] = C[i_c, j_c] = \sum_{\tau_1=0}^{i_a-1} \sum_{\tau_2=0}^{j_a-1} A[\tau_1, \tau_2] \times B[i - \tau_1, j - \tau_2]. \quad (2)$$

En conclusión, gran parte de la tecnología que existe hoy en día no sería posible sin un medio para extraer información y manipular señales digitales, donde uno de los métodos para procesar imágenes digitales, llamado filtrado, se puede utilizar para reducir la información de la señal no deseada (ruido) o extraer información como los bordes de la imagen, lo cual se logra mediante un enfoque matemático utilizando la convolución y la matriz del kernel.

El código empleado para el tratamiento de las imágenes en el software R versión 4.0.2 [6] se encuentra en el repositorio GitHub [1].

2. Aplicación de la prueba de chi cuadrada (χ^2)

Chi cuadrada (χ^2) es la distribución de la suma de variables aleatorias cuadradas. Se utiliza la distribución χ^2 para examinar si un conjunto de datos **difiere** de forma estadísticamente significativa de lo esperado, se conoce también como una prueba de calidad de ajuste y se requiere conocer la distribución que se espera ver y las frecuencias observadas de cada valor posible [5]. Se utiliza la ecuación 3 para determinar el valor de χ^2 , donde $k - 1$ son grados de libertad:

$$\sum_k \frac{(\text{esperada-observada})^2}{\text{esperada}} \sim \chi^2(|k| - 1). \quad (3)$$

En el tema de tesis que actualmente se desarrolla, se utiliza un GRASP (abreviatura de *greedy randomized adaptive search procedures*) reactivo para hallar la solución al problema. Este algoritmo depende de

Cuadro 1: Tabla de contingencia para la prueba χ^2

Alfa	Frecuencia		Diferencia cuadrada		
	Observada (E)	Esperada (E)	E-O	$(E - O)^2$	Normalizada
0.06	105	100	-5	25	0.25
0.07	99	100	1	1	0.01
0.08	113	100	-13	169	1.69
0.09	124	100	-24	576	5.76
0.10	110	100	-10	100	1
0.15	90	100	10	100	1
0.20	106	100	-6	36	0.36
0.25	84	100	16	256	2.56
0.30	82	100	18	324	3.24
0.35	87	100	13	169	1.69
Total	1,000	1,000	0	1,756	17.56

un valor alfa (α) para construir la solución. Para la afinación del parámetro α se decidió utilizar una estrategia reactiva en la cual el valor de α se adapta dinámicamente según los resultados obtenidos en las iteraciones previas, por lo tanto, se utiliza un conjunto discreto de valores predeterminados de alfas $A = \{\alpha_1, \alpha_2, \dots, \alpha_k\}$.

En la experimentación llevada a cabo para la tesis, se tiene en total de $k = 10$ valores de alfa diferentes y se ejecutaron $n = 1,000$ iteraciones del algoritmo para cada instancia, actualizando la probabilidad de selección para cada alfa cada 100 iteraciones. Para la aplicación de la prueba χ^2 se plantea como hipótesis nula (H_0) que las alfas en cada instancia son seleccionadas con la misma probabilidad y, como hipótesis alternativa (H_1) que las alfas en cada instancia no son seleccionadas con la misma probabilidad.

En el cuadro 1, se muestra la tabla de contingencia para la aplicación de la prueba α . La frecuencia de cada alfa se tomó de los resultados arrojados por el algoritmo y el valor esperado de cada alfa se determinó mediante la ecuación 4:

$$\text{Frecuencia esperada} = P(\alpha) \times n \quad (4)$$

$$\text{Frecuencia esperada} = \frac{1}{10} \times 1000 \quad (5)$$

$$\text{Frecuencia esperada} = 100.$$

Para ejecutar la prueba de χ^2 se utilizó la función `chisq.test()` del programa R, obteniendo como resultado un $\chi^2 = 17.56$ con 9 grados de libertad y un valor $p = 0.040$, por lo cual se procede a *rechazar* la hipótesis nula con un nivel de confianza del 95 %, lo que implica que al parecer se está seleccionando con mayor frecuencia algún valor de alfa dentro del conjunto dado. Este resultado tiene mucha lógica, ya que el algoritmo va seleccionando con mayor frecuencia aquellos valores de alfa que presentan mejor desempeño. El código empleado en el software R versión 4.0.2 para la prueba de chi cuadrada (χ^2) se encuentra en el repositorio GitHub [1].

```
test.txt
```

```
Chi-squared test for given probabilities

data: tabla
X-squared = 17.56, df = 9, p-value = 0.04064
```

3. Demostraciones numérica y analítica

Sea X, Y variables aleatorias y a, b, c, d son constantes conocidas, demostrar que:

- (a) $\text{Cov}(aX + b, cY + d) = ac\text{Cov}[X, Y]$.
- (b) $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$.

Con base en el desarrollo del teorema 6.2 y 6.4 del libro de Grinstead [2], se tiene que:

$$\begin{aligned}
 \text{Cov}(aX + b, cY + d) &= E[(aX + b)(cY + d)] - E[(aX + b)]E[(cY + d)] \\
 &= E(acXY + aXd + bcY + bd)[aE(X) + b][cE(Y) + d] \\
 &= acE(XY) + adE(X) + bcE(Y) + bd - [acE(X)E(Y) + adE(X) + bcE(Y) + bd] \\
 &= acE(XY) + adE(X) + bcE(Y) + bd - acE(X)E(Y) - adE(X) - bcE(Y) - bd \\
 &= acE(XY)acE(X)E(Y) \\
 &= ac\text{Cov}(X, Y).
 \end{aligned} \tag{6}$$

De acuerdo al desarrollo del teorema 6.6 y 6.8 del libro de Grinstead [2], se tiene que:

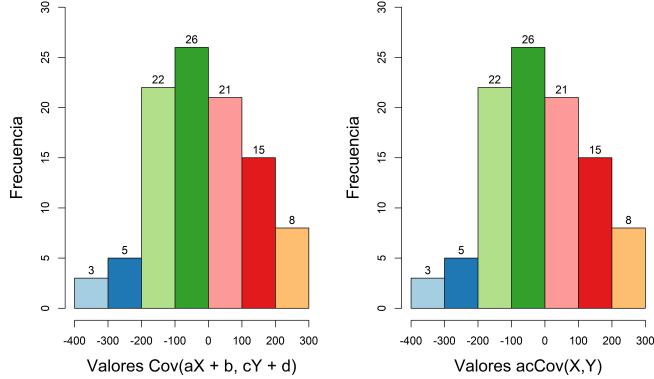
$$\begin{aligned}\text{Var}[X + Y] &= E[(X + Y)^2][E(X) + E(Y)]^2 \\&= E(X^2 + 2XY + Y^2) - [E(X)^2 + 2E(X)E(Y) + E(Y)^2] \\&= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\&= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 + 2[E(XY) - E(X)E(Y)] \\&= \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y].\end{aligned}\tag{7}$$

Para la demostración numérica, se generaron 100 valores aleatorios para X y Y (donde no necesariamente son independientes) y se estableció $a = 5$, $b = 100$, $c = 20.5$ y $d = 0.85$. Se realizaron 100 replicas aplicando las ecuaciones 6 y 7, donde X y Y tenían diferente tipo de distribución (uniforme, normal, exponencial y Poisson). Los resultados de la simulación para la covarianza y varianza de estos valores se muestran en las figuras 3 y 4, respectivamente, en las que se puede observar que se cumple la igualdad de las ecuaciones 6 y 7, respectivamente, sin importar el tipo de distribución y o la independencia de variables aleatorias X y Y (ver figuras 3c y 4c).

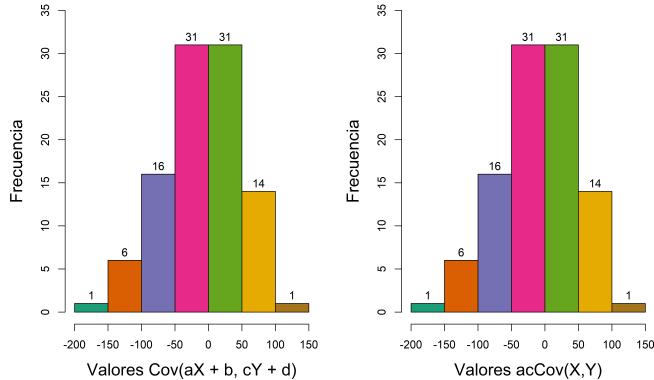
Todas las simulaciones fueron realizadas en el software R versión 4.0.2 y el código empleado se encuentra en el repositorio GitHub [1].

Referencias

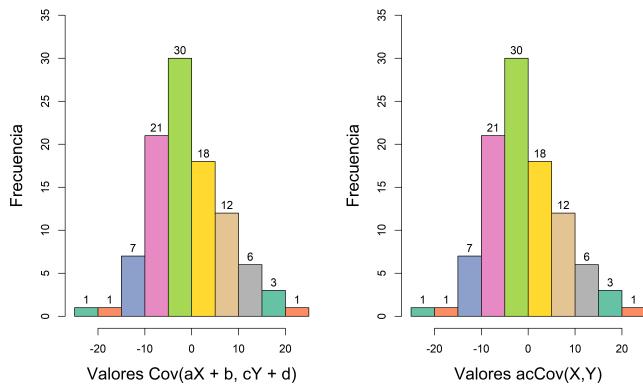
- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea11, 2020.
- [2] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability*. American Mathematical Society, 2006.
- [3] Kim, S., Casper, R. Applications of Convolution in Image Processing with MATLAB. *University of Washington*, 2013.
- [4] Mac Developer Library. Performing Convolution Operations. Recurso disponible en, <https://developer.apple.com/library/archive/documentation/Performance/Conceptual/vImage/ConvolutionOperations.html> 2016.
- [5] Schaeffer, Elisa. Modelos probabilistas aplicados: notas del curso. Recurso disponible en, [https://elisa.dyndns-web.com/teaching/prob/pasisis/prob.htmlut2](http://elisa.dyndns-web.com/teaching/prob/pasisis/prob.htmlut2).
- [6] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [7] Thyssen, Anthony. ImageMagick v6 Examples – Convolution of Images. Recurso disponible en, <https://legacy.imagemagick.org/Usage/convolve/>, 2013.



(a) Con $X \sim N(\mu, \sigma)$ y $Y \sim \text{Pois}(\lambda)$

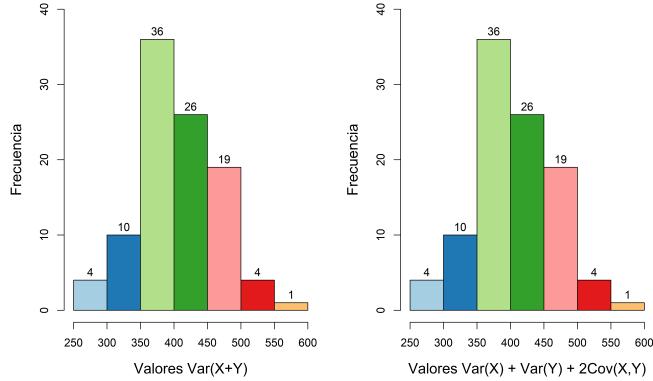


(b) Con $X \sim U(0, 1)$ y $Y \sim \text{Exp}(\lambda)$

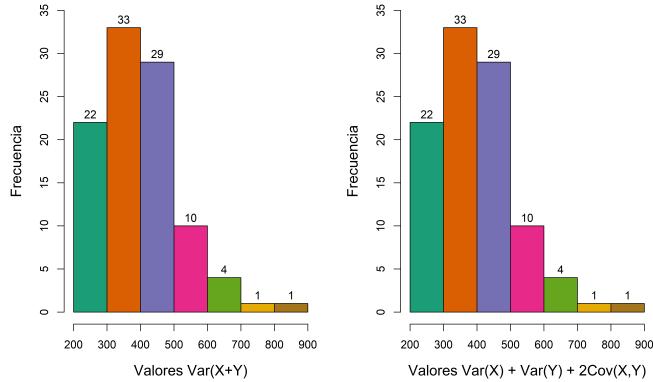


(c) Con $X \sim \text{Exp}(\lambda)$ y $Y \sim \text{Pois}(\lambda) + X$

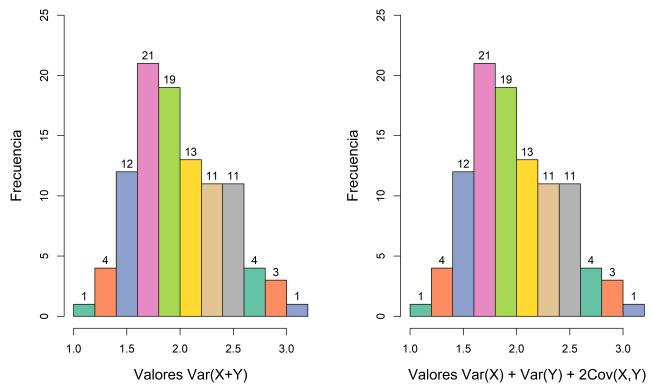
Figura 3: Resultados simulación para la covarianza



(a) Con $X \sim N(\mu, \sigma)$ y $Y \sim Pois(\lambda)$



(b) Con $X \sim U(0, 1)$ y $Y \sim Exp(\lambda)$



(c) Con $X \sim Exp(\lambda)$ y $Y \sim Pois(\lambda) + X$

Figura 4: Resultados simulación para la varianza

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 12

1. Problemas a resolver

En el presente trabajo se realizaron las soluciones de diversos problemas del libro de Grinstead [2] sobre las funciones generadoras de momentos de una variable aleatoria (v.a), así como el valor esperado y varianza. Para el cálculo de estos valores se utilizaron las ecuaciones 1, 2, 3, donde X es una v.a con distribución de probabilidad $f(x)$:

Función generadora de momentos si X es continua,

$$g(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} f(x) dx. \quad (1)$$

Valor esperado si X es continua,

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx. \quad (2)$$

Varianza de X ,

$$\sigma^2 = V(X) = E(X^2) - \mu^2 \quad (3)$$

1.1. Problema 1, página 392

Let Z_1, Z_2, \dots, Z_N describe a branching process in which each parent has j offspring with probability p_j .

Find the probability d that the process eventually dies out if:

a) $p_0 = 1/2, p_1 = 1/4, p_2 = 1/4$

b) $p_0 = 1/3, p_1 = 1/3, p_2 = 1/3$

c) $p_0 = 1/3, p_1 = 0, p_2 = 2/3$

d) $p_j = 1/2^{j+1}$

e) $p_j = (1/3)(2/3)^j$

f) $p_j = \frac{e^{-2}2^j}{j!}$

De acuerdo con el teorema 10.2, se tiene que si $m \leq 1$, entonces $d = 1$ y el proceso acaba con probabilidad 1; si $m > 1$, entonces $d < 1$ y el proceso acaba con probabilidad d . Para el cálculo del valor de m se utiliza las siguientes expresiones:

$$m = p_1 + 2p_2 = 1 - p_0 - p_2 + 2p_2 = 1 - p_0 + p_2$$

$$h(z) = p_0 + p_1 z + p_2 z^2 + \dots$$

$$m = h'(1).$$

Inciso a)

$$\begin{aligned} m &= \frac{1}{4} + 2\left(\frac{1}{4}\right) \\ m &= \frac{3}{4}. \end{aligned}$$

Como $m \leq 1$ y $p_0 > p_2$, entonces $d = 1$ y el proceso acaba con una probabilidad de 1.

Inciso b)

$$\begin{aligned} m &= \frac{1}{3} + 2\left(\frac{1}{3}\right) \\ m &= 1. \end{aligned}$$

Como $m \leq 1$ y $p_0 = p_2$, entonces $d = 1$ y el proceso acaba con una probabilidad de 1.

Inciso c)

$$\begin{aligned} m &= 0 + 2\left(\frac{2}{3}\right) \\ m &= \frac{4}{3}. \end{aligned}$$

Como $m > 1$ y $p_0 < p_2$, entonces $d < 1$ y el proceso acaba con una probabilidad de d . Para el cálculo del valor de d , se utiliza la ecuación 4:

$$\begin{aligned} d &= \frac{p_0}{p_2} & (4) \\ d &= \frac{\frac{1}{3}}{\frac{2}{3}} \\ d &= \frac{1}{2} \end{aligned}$$

Inciso d)

$$\begin{aligned}
h(z) &= \frac{1}{2^{0+1}} + \frac{1}{2^{1+1}}z + \frac{1}{2^{2+1}}z^2 + \dots \\
&= \frac{1}{2^1} + \frac{1}{2^2}z + \frac{1}{2^3}z^2 + \dots \\
&= \frac{1}{2} \left(1 + \frac{1}{2^1}z + \frac{1}{2^2}z^2 + \dots \right) \\
&= \frac{1}{2} \left(\frac{1}{1 - \frac{1}{2}z} \right) \\
&= \frac{1}{2-z} \\
h'(z) &= -\frac{\frac{d}{dz}(2-z)}{(2-z)^2} \\
&= -\frac{0-1}{(2-z)^2} \\
&= \frac{1}{(2-z)^2} \\
m = h'(1) &= \frac{1}{(2-1)^2} \\
&= 1
\end{aligned}$$

Como $m \leq 1$, entonces $d = 1$ y el proceso acaba con una probabilidad de 1.

Inciso e)

$$\begin{aligned}
h(z) &= \left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^0 + \left(\left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^1\right) z + \left(\left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^2\right) z^2 + \dots \\
&= \left(\frac{1}{3}\right) + \left(\left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^1\right) z + \left(\left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^2\right) z^2 + \dots \\
&= \frac{1}{3} \left(1 + \left(\frac{2}{3}\right)^1 z + \left(\frac{2}{3}\right)^2 z^2 + \dots \right) \\
&= \frac{1}{3} \left(\frac{1}{1 - \frac{2}{3}z} \right) \\
&= \frac{1}{3-2z} \\
h'(z) &= -\frac{\frac{d}{dz}(2-z)}{(2-z)^2} \\
&= -\frac{\frac{d}{dz}(3-2z)}{(3-2z)^2} \\
&= -\frac{0-2}{(3-2z)^2} \\
&= \frac{2}{(3-2z)^2} \\
m = h'(1) &= \frac{2}{(3-2)^2} \\
&= 2
\end{aligned}$$

Como $m > 1$, entonces $d < 1$ y el proceso acaba con una probabilidad de d . Para el cálculo del valor de d , se utiliza la ecuación 5:

$$\begin{aligned} z &= h(z) \\ &= \frac{1}{3 - 2z} \\ z(3 - 2z) &= 1 \\ 2z^2 - 3z + 1 &= 0 \end{aligned} \tag{5}$$

Por lo tanto, resolviendo la ecuación cuadrática queda que $z_1 = 1$ y $z_2 = 1/2 = d$.

Inciso f)

Considerando que $d = h(d) = p_0 + p_1d + p_2d^2 + \dots$, se utiliza esta expresión para estimar el valor de d numéricamente en el programa R, lo cual da como resultado $d \approx 0.2032$. El código empleado se encuentra en el repositorio GitHub [1].

1.2. Problema 3, página 401

In the chain letter problem (see Example 10.14) find your expected profit if:

a) $p_0 = 1/2, p_1 = 0, p_2 = 1/2$

b) $p_0 = 1/2, p_1 = 0, p_2 = 1/2$

Show that if $p_0 > 1/2$, you cannot expect to make a profit.

Solución

En este problema se tiene que el número esperado de cartas que se pueden vender es $m=p_1 + 2p_2$ y el valor esperado de la ganancia es $E_{(\text{ganancia})} = 50(m + m^2) - 100$, entonces se tiene que:

Inciso a)

$$m = 0 + 2 \left(\frac{1}{2} \right)$$

$$m = 1.$$

$$E_{(\text{ganancia})} = 50(1 + 1^{12}) - 100$$

$$E_{(\text{ganancia})} = 0.$$

Inciso b)

$$\begin{aligned}m &= \frac{1}{2} + 2\left(\frac{1}{3}\right) \\m &= \frac{7}{6}. \\E_{(\text{ganancia})} &= 50\left(\frac{7}{6} + \left(\frac{7}{6}\right)^{12}\right) - 100 \\E_{(\text{ganancia})} &\approx 276.26.\end{aligned}$$

Demostración

Se considera que $p_0 + p_1 + p_2 = 1$, entonces se contempla un $p_0 = 0.55$, $p_1 = 0.25$ y $p_2 = 0.2$, y se obtiene lo siguiente:

$$\begin{aligned}m &= 0.25 + 2(0.2) \\m &= 0.65 \\E_{(\text{ganancia})} &= 50(0.65 + 0.65^{12}) - 100 \\E_{(\text{ganancia})} &\approx -67.22.\end{aligned}$$

1.3. Problema 1, página 401

Let X be a continuous random variable with values in $[0, 2]$ and density f_X . Find the moment generating function $g(t)$ for X if:

a) $f_X(x) = 1/2$

$$\begin{aligned}g(t) &= \int_0^2 e^{xt} \frac{1}{2} dx \longrightarrow (\text{ecuación 1}) \\&= \frac{1}{2} \int_0^2 e^{xt} dx \\&= \frac{1}{2} \left[\frac{e^{xt}}{t} \Big|_0 \right] \\&= \frac{1}{2} \left[\frac{e^{2t}}{t} - \frac{e^{0t}}{t} \right] \\&= \frac{1}{2} \left[\frac{e^{2t} - 1}{t} \right] \\&= \frac{e^{2t} - 1}{2t}. \tag{6}\end{aligned}$$

b) $f_X(x) = (1/2)x$

$$\begin{aligned}
g(t) &= \int_0^2 e^{xt} \frac{1}{2}x dx \longrightarrow (\text{ecuación 1}) \\
&= \frac{1}{2} \int_0^2 e^{xt} x dx \implies UV - \int_0^2 V dU, \quad U = x, \quad dU = dx, \quad V = \frac{e^{xt}}{t} \\
&= \frac{1}{2} \left[x \frac{e^{xt}}{t} - \int_0^2 \frac{e^{xt}}{t} dx \right] \\
&= \frac{1}{2} \left[x \frac{e^{xt}}{t} - \frac{1}{t} \int_0^2 e^{xt} dx \right] \\
&= \frac{1}{2} \left[x \frac{e^{xt}}{t} - \frac{e^{xt}}{t^2} \Big|_0 \right] \\
&= \frac{1}{2} \left[2 \frac{e^{2t}}{t} - \frac{e^{2t}}{t^2} - \left(0 \frac{e^{0t}}{t} - \frac{e^{0t}}{t^2} \right) \right] \\
&= \frac{1}{2} \left[\frac{2e^{2t}}{t} - \frac{e^{2t}}{t^2} + \frac{1}{t^2} \right] \\
&= \frac{1}{2} \left[\frac{2te^{2t} - e^{2t} + 1}{t^2} \right] \\
&= \frac{e^{2t}(2t - 1) + 1}{2t^2}. \tag{7}
\end{aligned}$$

c) $f_X(x) = 1 - (1/2)x$

$$\begin{aligned}
g(t) &= \int_0^2 e^{xt} \left[1 - \frac{1}{2}x \right] dx \longrightarrow (\text{ecuación 1}) \\
&= \int_0^2 e^{xt} - \frac{e^{xt}}{2}x dx \\
&= \int_0^2 e^{xt} dx - \int_0^2 \frac{e^{xt}}{2}x dx \\
&= \int_0^2 e^{xt} dx - \frac{1}{2} \int_0^2 e^{xt} x dx \\
&= \frac{\cancel{e^{2t}} - 1}{\cancel{t}} - \frac{e^{2t}(2t - 1) + 1}{2t^2} \longrightarrow (\text{ecuación 6 y ecuación 8}) \\
&= \frac{2e^{2t} - 2t - 2e^{2t} + e^{2t} - 1}{2t^2} \\
&= \frac{e^{2t} - 2t - 1}{2t^2}.
\end{aligned}$$

d) $f_X(x) = |1 - x|$

$$\begin{aligned}
g(t) &= \int_0^2 e^{xt}(|1-x|) dx \longrightarrow (\text{ecuación 1}) \\
&= \int_0^1 e^{xt}(|1-x|) dx + \int_1^2 e^{xt}(|x-1|) dx \\
&= \int_0^1 e^{xt} - e^{xt}x dx + \int_1^2 e^{xt}x - e^{xt} dx \\
&= \left[\int_0^1 e^{xt} - \int_0^1 e^{xt}x dx \right] + \left[\int_1^2 e^{xt}x - \int_1^2 e^{xt} dx \right] \\
&= \left[\left[\frac{e^{xt}}{t} - \left(\frac{xe^{xt}}{t} - \frac{e^{xt}}{t^2} \right) \right] \Big|_0^1 \right] + \left[\left[\frac{xe^{xt}}{t} - \frac{e^{xt}}{t^2} - \frac{e^{xt}}{t} \right] \Big|_1^2 \right] \longrightarrow (\text{ecuación 6 y ecuación 7}) \\
&= \left[\left[\frac{e^{xt}}{t} - \frac{xe^{xt}}{t} + \frac{e^{xt}}{t^2} \right] \Big|_0^1 \right] + \left[\left[\frac{xe^{xt}}{t} - \frac{e^{xt}}{t^2} - \frac{e^{xt}}{t} \right] \Big|_1^2 \right] \\
&= \left[\left[\frac{e^{xt} - xe^{xt}}{t} + \frac{e^{xt}}{t^2} \right] \Big|_0^1 \right] + \left[\left[\frac{xe^{xt}}{t} - \frac{e^{xt}}{t^2} - \frac{e^{xt}}{t} \right] \Big|_1^2 \right] \\
&= \left[\frac{e^{1t} - 1e^{1t}}{t} + \frac{e^{1t}}{t^2} - \left(\frac{e^{0t} - 0e^{0t}}{t} + \frac{e^{0t}}{t^2} \right) \right] + \left[\frac{2e^{2t}}{t} - \frac{e^{2t}}{t^2} - \frac{e^{2t}}{t} - \left(\frac{1e^{1t}}{t} - \frac{e^{1t}}{t^2} - \frac{e^{1t}}{t} \right) \right] \\
&= \left[\frac{e^t}{t^2} - \frac{1}{t} - \frac{1}{t^2} \right] + \left[\frac{e^{2t}}{t} - \frac{e^{2t}}{t^2} + \frac{e^t}{t^2} \right] \\
&= \left[\frac{e^t - t - 1}{t^2} \right] + \left[\frac{te^{2t} - e^{2t} + e^t}{t^2} \right] \\
&= \frac{te^{2t} - e^{2t} + 2e^t - t - 1}{t^2} \\
&= \frac{e^{2t}(t-1) + 2e^t - t - 1}{t^2}.
\end{aligned}$$

e) $f_X(x) = (3/8)x^2$

$$\begin{aligned}
g(t) &= \int_0^2 e^{xt} \frac{3}{8} x^2 dx \longrightarrow (\text{ecuación 1}) \\
&= \int_0^2 \frac{3x^2 e^{xt}}{8} dx \\
&= \frac{3}{8} \int_0^2 x^2 e^{xt} dx \implies UV - \int_0^2 V dU, \quad U = x^2, \quad dU = 2x dx, \quad V = \frac{e^{xt}}{t} \\
&= \frac{3}{8} \left[x^2 \frac{e^{xt}}{t} - \int_0^2 \frac{e^{xt}}{t} 2x dx \right] \\
&= \left[x^2 \frac{e^{xt}}{t} - \frac{2}{t} \int_0^2 e^{xt} x dx \right] \\
&= \frac{3}{8} \left[\left[x^2 \frac{e^{xt}}{t} - \frac{2}{t} \left(x \frac{e^{xt}}{t} - \frac{e^{xt}}{t^2} \right) \right] \Big|_0^2 \right] \\
&= \frac{3}{8} \left[\left[x^2 \frac{e^{xt}}{t} - \frac{2}{t} \left(\cancel{x} \frac{e^{xt}}{t} - \frac{e^{xt}}{\cancel{t}^2} \right) \right] \Big|_0^2 \right] \longrightarrow (\text{ecuación 7}) \\
&= \frac{3}{8} \left[\left[\frac{x^2 e^{xt}}{t} - \frac{2}{t} \left(\frac{tx e^{xt}}{t^2} - \frac{e^{xt}}{t^2} \right) \right] \Big|_0^2 \right] \\
&= \frac{3}{8} \left[\left[\frac{x^2 e^{xt}}{t} - \frac{2tx e^{xt} + 2e^{xt}}{t^3} \right] \Big|_0^2 \right] \\
&= \frac{3}{8} \left[\left[\frac{t^2 x^2 e^{xt} - 2tx e^{xt} + 2e^{xt}}{t^3} \right] \Big|_0^2 \right] \\
&= \frac{3}{8} \left[\frac{t^2(2)^2 e^{2t} - 2t(2)e^{2t} + 2e^{2t}}{t^3} - \left(\frac{t^2(0)^2 e^{0t} - 2t(0)e^{0t} + 2e^{0t}}{t^3} \right) \right] \\
&= \frac{3}{8} \left[\frac{4t^2 e^{2t} - 4te^{2t} + 2e^{2t} - 2}{t^3} \right] \\
&= \frac{12t^2 e^{2t} - 12te^{2t} + 6e^{2t} - 6}{8t^3} \\
&= \frac{6t^2 e^{2t} - 6te^{2t} + 3e^{2t} - 3}{4t^3} \\
&= \frac{3}{4} \left[\frac{e^{2t}(2t^2 - 2t + 1) - 1}{t^3} \right].
\end{aligned}$$

1.4. Problema 6, página 402

Let X be a continuous random variable whose characteristic function $k - X(\tau)$ is:

$$k_X(\tau) = e^{-|\tau|} \quad -\infty < \tau < \infty. \quad (9)$$

Show directly that the density f_X of X is

$$f_X(x) = \frac{1}{\pi(1+x^2)}. \quad (10)$$

Solución

Se tiene que la función de densidad $f_X(x)$ es:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ix\tau} k_X(\tau) d\tau \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ix\tau} e^{-|\tau|} d\tau \longrightarrow (\text{ecuación 9}) \\ &= \frac{1}{2\pi} \left[\int_{-\infty}^0 e^{-ix\tau} e^{-(-\tau)} d\tau + \int_0^{\infty} e^{-ix\tau} e^{-\tau} d\tau \right] \\ &= \frac{1}{2\pi} \left[\int_{-\infty}^0 e^{-ix\tau+\tau} d\tau + \int_0^{\infty} e^{-ix\tau-\tau} d\tau \right] \\ &= \frac{1}{2\pi} \left[\int_{-\infty}^0 e^{\tau(1-ix)} d\tau + \int_0^{\infty} e^{-\tau(ix+1)} d\tau \right] \\ &= \frac{1}{2\pi} \left[\left[\frac{e^{\tau(1-ix)}}{(1-ix)} \right] \Big|_{-\infty}^0 + \left[-\frac{e^{-\tau(ix+1)}}{(ix+1)} \right] \Big|_0^{\infty} \right] \\ &= \frac{1}{2\pi} \left[\left[\frac{e^{0(1-ix)}}{(1-ix)} - \frac{e^{-\infty(1-ix)}}{(1-ix)} \right] - \left[\frac{e^{-\infty(ix+1)}}{(ix+1)} - \frac{e^{-0(ix+1)}}{(ix+1)} \right] \right] \\ &= \frac{1}{2\pi} \left[\frac{1}{(1-ix)} + \frac{1}{(ix+1)} \right] \\ &= \frac{1}{2\pi} \left[\frac{ix+1+1-ix}{(ix+1-(ix)^2-ix)} \right] \\ &= \frac{1}{2\pi} \left[\frac{2}{(1-(ix)^2)} \right] \\ &= \frac{1}{\pi} \left[\frac{1}{(1-(-x^2))} \right] \longrightarrow (\textcolor{blue}{i = \sqrt{-1}}, (\text{teorema 10.4})) \\ &= \frac{1}{\pi(1+x^2)} \longrightarrow (\text{ecuación 10}). \end{aligned}$$

1.5. Problema 10, página 403

Let X_1, X_2, \dots, X_n , be an independent trials process with density

$$f(x) = \frac{1}{2} e^{-|x|} \quad -\infty < x < \infty.$$

a) Find the mean and variance of $f(x)$

$$\begin{aligned}\mu = E(X) &= \int_{-\infty}^{\infty} x \frac{1}{2} e^{-|x|} dx \longrightarrow (\text{ecuación 2}) \\ &= \frac{1}{2} \left[\int_{-\infty}^0 x e^{-(x)} dx + \int_0^{\infty} x e^{-(x)} dx \right] \\ &= \frac{1}{2} \left[\int_{-\infty}^0 x e^{(x)} dx + \int_0^{\infty} x e^{-(x)} dx \right] \\ &= \frac{1}{2} [-1 + 1] \\ &= 0.\end{aligned}$$

Ahora, para calcular la varianza, se necesita el valor de $E(X^2)$, entonces:

$$\begin{aligned}\mu = E(X) &= \int_{-\infty}^{\infty} x^2 \frac{1}{2} e^{-|x|} dx \longrightarrow (\text{ecuación 2}) \\ &= \frac{1}{2} \left[\int_{-\infty}^0 x^2 e^{-(x)} dx + \int_0^{\infty} x^2 e^{-(x)} dx \right] \\ &= \frac{1}{2} \left[\int_{-\infty}^0 x^2 e^{(x)} dx + \int_0^{\infty} x^2 e^{-(x)} dx \right] \\ &= \frac{1}{2} [2 + 2] \\ &= 2.\end{aligned}$$

Por lo tanto, para calcular la varianza se utiliza la ecuación 3:

$$V(X) = 2 - (0)^2$$

$$= 2 - 0$$

$$= 2.$$

b) Find the moment generating function for X_1 , S_n , A_n , and S_n^*

$$\begin{aligned}
g(t) &= \int_{-\infty}^{\infty} e^{xt} \left[\frac{e^{-|x|}}{2} \right] dx \longrightarrow (\text{ecuación 1}) \\
&= \frac{1}{2} \int_{-\infty}^{\infty} e^{xt} e^{-|x|} dx \\
&= \frac{1}{2} \left[\int_{-\infty}^0 e^{xt} e^{-(x)} dx + \int_0^{\infty} e^{xt} e^{-(x)} dx \right] \\
&= \frac{1}{2} \left[\int_{-\infty}^0 e^{xt+x} dx + \int_0^{\infty} e^{xt-x} dx \right] \\
&= \frac{1}{2} \left[\int_{-\infty}^0 e^{x(t+1)} dx + \int_0^{\infty} e^{-x(-t+1)} dx \right] \\
&= \frac{1}{2} \left[\left[\frac{e^{x(t+1)}}{(t+1)} \right] \Big|_{-\infty}^0 + \left[-\frac{e^{-x(1-t)}}{(1-t)} \right] \Big|_0^{\infty} \right] \\
&= \frac{1}{2} \left[\left[\frac{e^{0(t+1)}}{(t+1)} - \frac{e^{-\infty(t+1)}}{t+1} \right] - \left[\frac{e^{-\infty(1-t)}}{(1-t)} - \frac{e^{-0(1-t)}}{(1-t)} \right] \right] \\
&= \frac{1}{2} \left[\frac{1}{(t+1)} + \frac{1}{(1-t)} \right] \\
&= \frac{1}{2} \left[\frac{1-t+t+1}{(t-t^2+1-t)} \right] \\
&= \frac{1}{2} \left[\frac{2}{(1-t^2)} \right] \\
&= \frac{1}{(1-t^2)}.
\end{aligned}$$

$$\begin{aligned}
S_n &= (g(t))^n \\
&= \left(\frac{1}{(1-t^2)} \right)^n \\
&= \frac{1}{(1-t^2)^n}.
\end{aligned}$$

$$\begin{aligned}
S_n^* &= \left(g\left(\frac{t}{\sqrt{n}}\right) \right)^n \\
&= \left(\frac{1}{\left(1 - \left(\frac{t}{\sqrt{n}} \right)^2 \right)} \right)^n \\
&= \frac{1}{\left(1 - \left(\frac{t}{\sqrt{n}} \right)^2 \right)^n}.
\end{aligned}$$

c) What can you say about the moment generating function of S_n^* as $n \rightarrow \infty$?

$$\lim_{n \rightarrow \infty} \frac{1}{\left(1 - \left(\frac{t}{\sqrt{n}}\right)^2\right)^n}$$

Por lo tanto, la función generadora de $S_n^* = 1$.

Referencias

- [1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea12, 2020.
- [2] Grinstead, Charles M., Snell, J. Laurie. *Introduction to Probability*. American Mathematical Society, 2006.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 13

1. Aplicación de la ley de números grandes

La ley de los números grandes indica que, en especial, el promedio de un gran número de resultados refleja o se acerca al valor esperado (media analítica) y que la diferencia se reduce a medida que se introducen más resultados. Por ejemplo, en la prueba de Bernoulli con probabilidad de éxito p : con suficientes repeticiones, el porcentaje de éxitos obtenidos se acerca necesariamente a p [5].

Existen dos formas de la ley de los grandes números, **ley débil** y **la ley fuerte**, en referencia a dos modos diferentes de convergencia del promedio de la muestra acumulada, entonces sean X_1, X_2, \dots, X_n un proceso de pruebas independientes e igualmente distribuidos (i.i.d) con un valor esperado finito $\mu = E[X]$ y una varianza finita $\sigma^2 = E[X^2] - E[X]^2$ y sean $S_n = X_1 + X_2 + \dots + X_n$ y ϵ un valor real arbitrario positivo, **la ley débil** (también llamada ley de *Khinchin*) establece que el promedio de la muestra converge hacia el valor esperado, es decir, que la probabilidad de que estos valores sean diferentes es cero y, **la ley fuerte** establece que el promedio de la muestra converge casi con seguridad al valor esperado, es decir, que la probabilidad de que estos valores sean iguales es uno [3]. Lo anterior se expresa mediante las ecuaciones 1 y 2, respectivamente:

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0 \quad \forall \epsilon > 0, \quad (1)$$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1 \quad \forall \epsilon > 0. \quad (2)$$

En la investigación desarrollada por Tinungki [8], utilizan la ley de los números en el campo de los seguros de vida. Aunque el seguro de vida es un negocio, solo lo es para aquellas empresas que pueden mantener su solidez financiera mientras pagan reclamaciones. Las compañías de seguros utilizan la ley

de los grandes números para reducir su propio riesgo de pérdida al agrupar un número suficientemente grande de personas en un grupo asegurado.

Cabe recordar que, en este negocio, los riesgos que enfrenta cada individuo son transferidos a la compañía de seguros, la cual se compromete a indemnizar el monto especificado en el contrato de la póliza. Para compensar esta pérdida, el asegurador fija la prima a pagar por el asegurado, por lo tanto, los errores en la medición de los factores que se involucran al establecer esta prima (el valor de cualquier pérdida, los costos administrativos, factores económicos, de salud y sociales, entre otros) pueden causar pérdidas a las compañías de seguros, en especial cuando se fija una prima menor de la que deberían. El seguro de vida, como herramienta para distribuir el riesgo, solo puede funcionar si una compañía puede asumir el mismo riesgo en grandes cantidades.

Para cuantificar el riesgo de fallecimiento de una determinada clase de riesgo, o para ser más exactos para estimar la cantidad de personas de determinada edad y determinada clase de riesgo que morirán cada año, las compañías de seguros de vida utilizan las tablas de mortalidad o nivel de morbilidad (el nivel de enfermedad, lesión y ocurrencia de fallas de salud). Estas tablas dan el porcentaje exacto de probabilidad de que alguien muera en un año determinado (media teórica) [4]. El uso de estas tasas permite que la compañía de seguros pueda predecir el potencial de pérdidas de sus clientes, a pesar de que estos datos pueden complicarse con el tiempo, se pueden predecir con precisión gracias a la ley de los grandes números, por lo tanto, el costo de las pérdidas se puede distribuir uniformemente sobre el total de clientes de acuerdo con la clase garantizada por el seguro. Por ende, con base en la definición de la ley de los grandes números, cuando se obtiene una muestra aleatoria tomada desde una variable aleatoria independientes e igualmente distribuida, con media y varianza finita, el promedio de la muestra estará cerca del promedio de la población [8]. Es así como el uso de esta ley permite predecir mejor el número de pérdidas, ya que a mayor sea la población asegurada, más precisas serán las predicciones de las perdidas esperadas y esto le permite a las aseguradoras fijar el precio de las pólizas de seguro con y cobro de la prima con precisión [6].

Se realizó una simulación como ejemplo numérico para demostrar lo anteriormente mencionado, es decir, a medida que aumenta la cantidad de personas aseguradas (tiende a ser igual la población total) se puede proyectar con mejor exactitud la perdidas u ocurrencia del evento que se está asegurando. Para esta simulación, se considera que cada año hay una probabilidad de $1/250$ de que ocurra un siniestro z , por lo tanto, nuestra variable aleatoria, en este caso variables booleanas independientes, tendrán el valor de 1 cuando ocurra el evento con probabilidad de $1/250$ (lo que representa una pérdida) y de 0 con probabilidad de $249/250$ ($1 - 1/250$), se calcula el promedio de la cantidad de ocurrencias del evento para 1,000 repeticiones y se va variando el tamaño de la muestra n (número de asegurados) desde 1,000, 100,000, 1,000,000 y 10,000,000 [1]. El código empleado para esta simulación se realizó en el programa R versión 4.0.2 [7] y se encuentra en el repositorio GitHub [2]. Agradecimientos al compañero Alberto

Benavidez por su colaboración en la implementación del código.

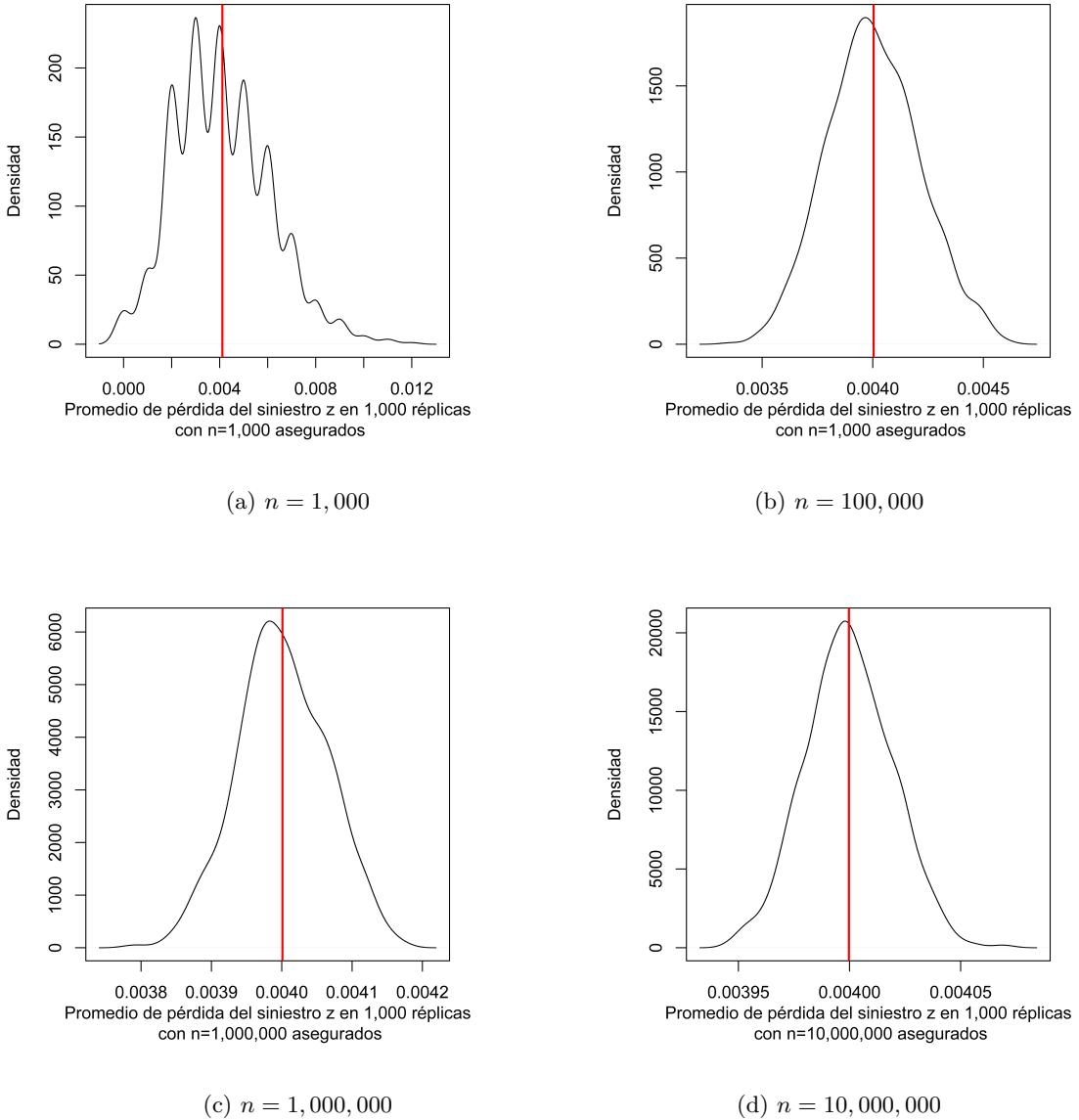


Figura 1: Resultados de la experimentación variando el n número de asegurados

En la figura 1 se muestran los resultados de la simulación en la que podemos observar que a medida que aumenta el número de riesgos independientes (es decir, el tamaño de la muestra crece), las probabilidades del número de pérdidas es cercano a la media esperada, por lo tanto más precisas serán las predicciones de las perdidas esperadas.

Referencias

- [1] Autor, David. Microeconomic Theory and Public Policy. Recurso disponible en, https://ocw.mit.edu/courses/economics/14-03-microeconomic-theory-and-public-policy-fall-2016/lecture-notes/MIT14_03F16_lec17.pdf, 2016.
- [2] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea13, 2020.
- [3] Loeve, M. *Probability Theory I*. Springer-Verlag New York, 4th edition, 1977.
- [4] Rockford, Thomas. What Is The Law Of Large Numbers? Recurso disponible en, <https://www.lifeant.com/what-is-the-law-of-large-numbers/>, 2019.
- [5] Schaeffer, Elisa. Modelos probabilistas aplicados: notas del curso. Recurso disponible en, <https://elisa.dyndns-web.com/teaching/prob/pisis/prob.html#t16>.
- [6] Smith, Michael L., Kane, Stephen A. *The Law of Large Numbers and the Strength of Insurance*, pages 1–27. Springer Netherlands, 1994.
- [7] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.
- [8] Tinungki, Georgina Maria. The Application Law of Large Numbers That Predicts The Amount of Actual Loss in Insurance of Life. *Journal of Physics: Conference Series*, 979, 2017.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 14

1. Aplicación del teorema de límite central

El teorema del límite central (CTL por sus siglas en inglés de, *Central Limit Theorem* es un teorema fundamental de probabilidad y estadística. Este teorema indica que cuando recolectamos una muestra suficientemente grande de n observaciones independientes de una población con media μ y varianza finita σ^2 , la distribución muestral de las medias muestrales sigue aproximadamente una distribución normal con media $= \mu$ y desviación estándar $= \sigma/\sqrt{n}$ [1].

Sean X_1, X_2, \dots, X_n un proceso de pruebas independientes e igualmente distribuidos (i.i.d) con media μ y una varianza finita $0 < \sigma^2 < \infty$ y sean $S_n = X_1 + X_2 + \dots + X_n$ y Φz la función de densidad de la distribución normal estándar, en la ecuación 1 se expresa una versión del teorema [5]:

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z). \quad (1)$$

Principales propiedades del teorema central del límite

De acuerdo a López [3] el teorema central del límite tiene una serie de propiedades de gran utilidad en el ámbito estadístico y probabilístico. Las principales son:

- Si el tamaño de la muestra es suficientemente grande, la distribución de las medias muestrales seguirá aproximadamente una distribución normal, lo cual se cumple independientemente de la forma de la distribución con la que se trabajando.
- En estadística, un tamaño de muestra lo suficientemente grande para sacar conclusiones es mayor o igual a 30.

-
- Selección al azar, para que no esté sesgado hacia ciertas características de la población.
 - La media poblacional y la media muestral serán iguales. Es decir, la media de la distribución de todas las medias muestrales será igual a la media del total de la población.

El teorema del límite central es una aproximación que se puede usar cuando la población que está estudiando es tan grande que tomaría mucho tiempo recopilar datos sobre cada individuo que forma parte de ella, por ende, en términos estadísticos, al recolectar muestras de una población en particular y combinar la información de las muestras, se podría sacar conclusiones sobre la población [1]. Además, cuando los datos tienden a tener una distribución normal este tipo de distribución es muy fácil de aplicar para realizar contrastes de hipótesis y construcción de intervalos de confianza [3].

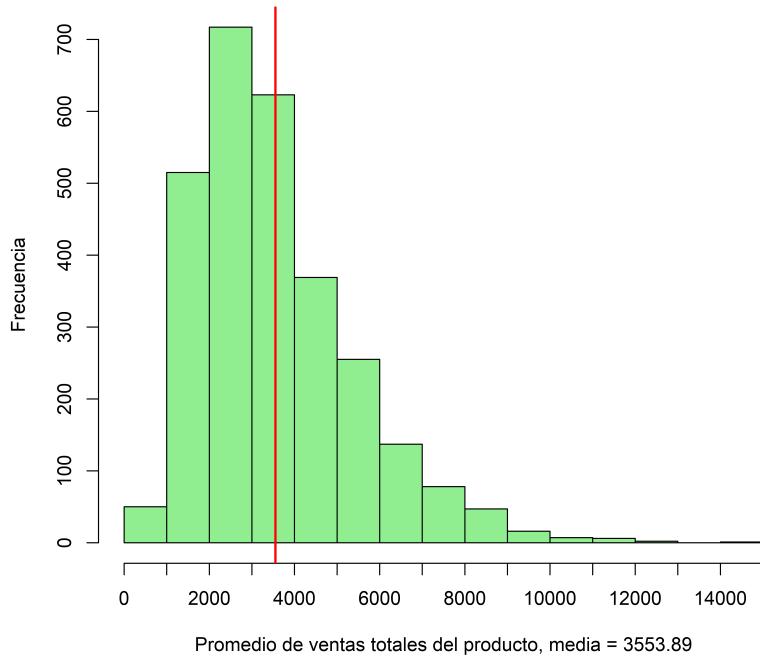
El caso aplicado de este teorema fue con base en el trabajo de Bento [1] donde se utiliza el CTL para determinar la cantidad de reabastecimiento semanal de un producto y no incurrir en exceso de inventario inactivo en las tiendas de una cadena de supermercados. Como se mencionó anteriormente, con base en el teorema del límite central, no es necesario tener que visitar todas las tiendas de la región y obtener las cifras de ventas del producto de la semana para saber cuántas unidades solicitar en el próximo pedido. Lo que se puede hacer es recopilar muchas muestras de las ventas semanales en las tiendas (la población), calcular su media (el número medio producto vendido) y construir la distribución de las medias de la muestra, conocida como la distribución muestral. Si estas muestras cumplen los criterios del CTL, se podrá que la distribución de las medias muestrales se puede aproximar a la distribución normal. Los criterios de la muestra serían los siguientes:

- Seleccionado al azar para evitar sesgo.
- Representativa de la población, mayor o igual a 30.
- Selección al azar, para que no esté sesgado hacia ciertas características de la población.
- Incluir menos del 10 % de la población, si se toman muestras sin reemplazo ya que las observaciones en la población no son todas independientes entre sí, si se recolecta una muestra que es demasiado grande, puede terminar recolectando observaciones que no son independientes entre sí. Incluso si esas observaciones se eligieron al azar.

Para la experimentación se utilizó una base de datos con datos de ventas, información de pedido, entre otros, de dominio público disponible en Kaggle [4]. Se realizó una simulación como ejemplo numérico para demostrar el CLT, es decir, con una muestra representativa de la población bajo los criterios mencionados anteriormente, la distribución de las medias muestrales a medida que aumenta el número de muestras tomadas, más se acerca a la forma de una distribución normal. Para esta experimentación, se calcula el promedio de ventas con 10, 100, 1,000 y 10,000 muestras de tamaños 30 y 100. El código

empleado para esta simulación se realizó en el programa R versión 4.0.2 [6] y junto con la base de datos utilizada se encuentran en el repositorio GitHub [2].

En la figura 1 se muestra el histograma de las ventas totales obtenidas, en la cual podemos observar que tienen una media aproximada de 5,553.89. En la figura 2 y 3 se muestran los resultados obtenidos para las 10, 100, 1,000 y 10,000 muestras de tamaño 30 y 100 respectivamente, en las cuales podemos observar que con 10,000 muestras aleatorias de tamaño 100 del conjunto de datos de ventas, se obtiene una distribución de muestreo que se asemeja a la curva de campana característica de la distribución normal con un promedio de ventas similar al de la población (ventas totales). Lo anterior sucede ya que, al recolectar una muestra más grande, se tendrá menos posibilidades de obtener valores extremos, por lo que sus valores estarán más agrupados y por lo tanto, la desviación estándar, o la distancia de la media, será menor. Otra forma de explicarlo sería considerando la ecuación 1, ya que la desviación estándar de la distribución muestral, también llamada error estándar, es igual a σ/\sqrt{n} , entonces, a medida que aumenta el tamaño de la muestra, el denominador también aumenta y hace que el valor estándar general sea más pequeño [1].



Referencias

- [1] Bento, Carolina. Central Limit Theorem: a real-life application. Recurso disponible en, <https://towardsdatascience.com/central-limit-theorem-a-real-life-application-f638657686e1>, 2015.
- [2] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea14, 2020.
- [3] López Abellán, Joaquín. Teorema central del límite (TCL). Recurso disponible en, <https://economipedia.com/definiciones/teorema-central-del-limite.html>, 2015.
- [4] Segura, Gus. Sample Sales Data. Recurso disponible en, <https://www.kaggle.com/kyanyoga/sample-sales-data>, 2016.
- [5] Stanton, Charles. The Central Limit Theorem. Recurso disponible en, <https://web.archive.org/web/20100602111757/http://www.math.csusb.edu/faculty/stanton/probstat/clt.html>, 2010.
- [6] The R Foundation. The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.

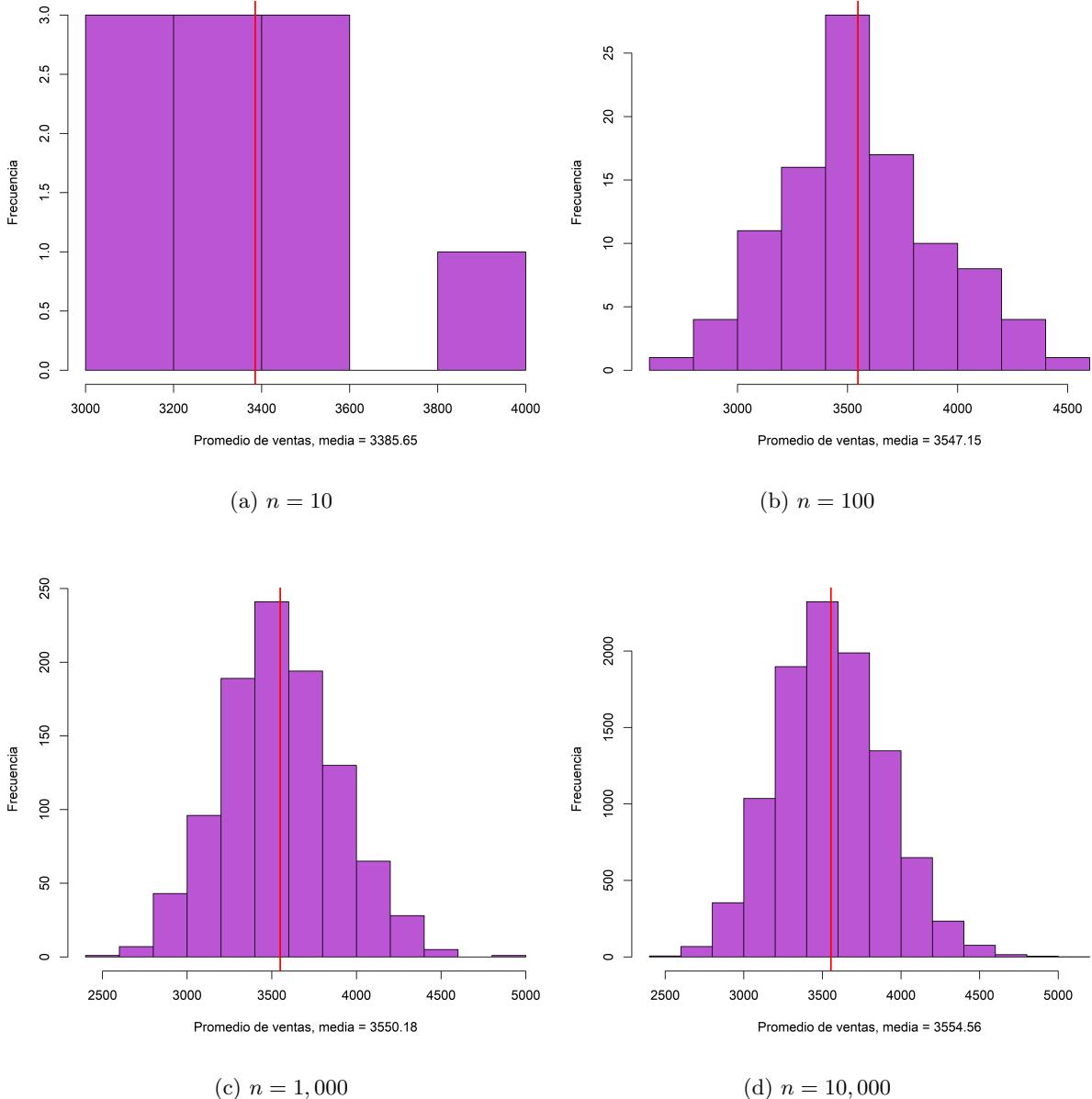


Figura 2: Resultados de la experimentación con n muestras de tamaño 30

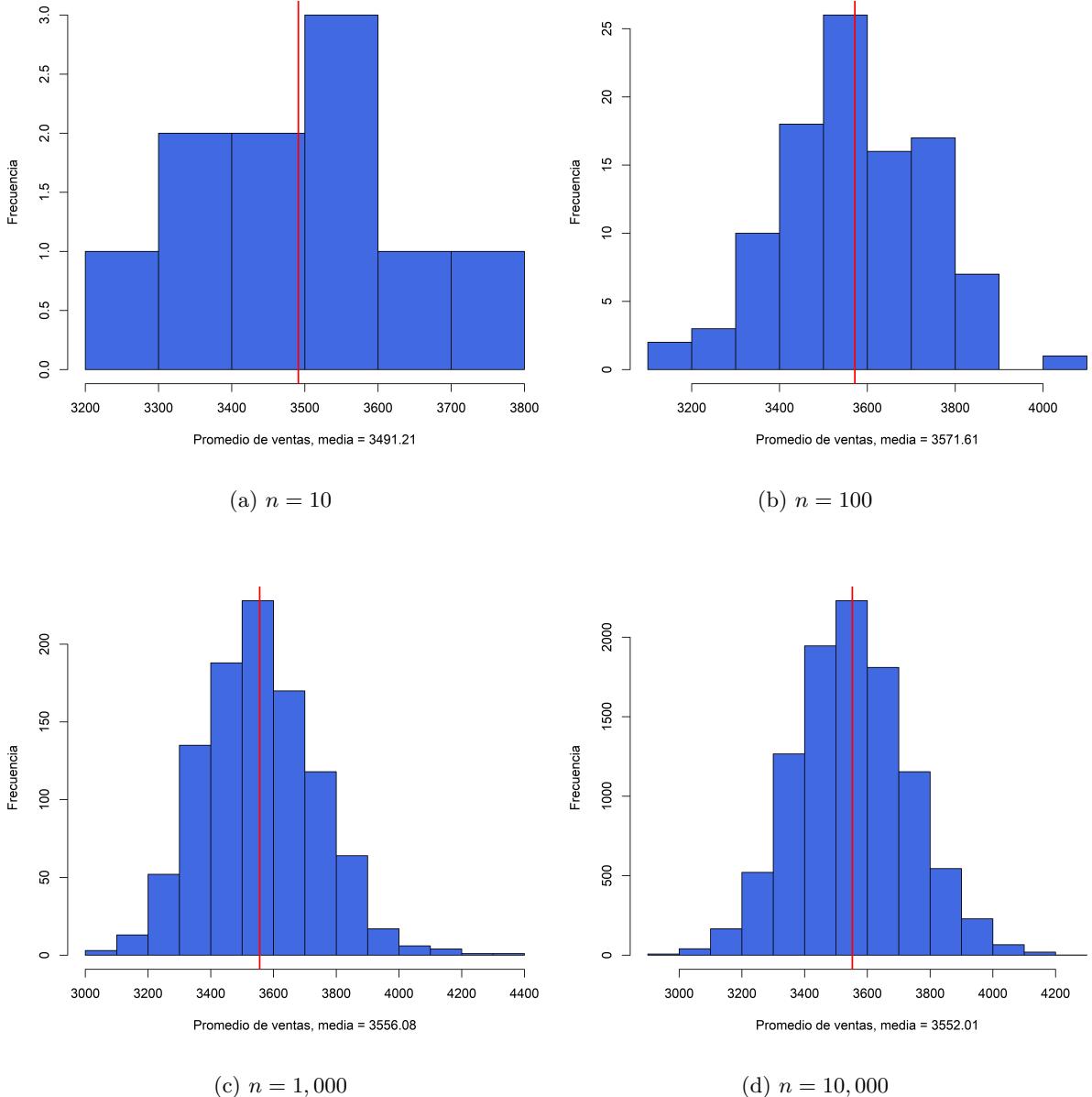


Figura 3: Resultados de la experimentación con n muestras de tamaño 100

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 15

1. Propuestas para el proyecto final de la clase

Las siguientes propuestas son planteadas como ideas para realizar el producto integrador de la materia Modelos Probabilistas Aplicados, las cuales podrían ser modificadas para llevar a cabo un trabajo final que cumpla con los requisitos:

1. En el tema de tesis, se cuenta con un modelo matemático y una metaheurística para encontrar la solución del problema planteado. Por medio de la prueba de hipótesis de medias de diferencia se pretende comprobar que existe un ahorro entre la metaheurística y la solución ofrecida por el modelo matemático en un 95 % y determinar que tanto mejora la solución considerando intervalos de confianza del 90 % y 95 %. Estas pruebas se aplicarán tanto por tamaño de instancias (pequeñas, medianas tipo 1, medianas tipo 2 y grandes) como para el total de instancias. Además, se aplicaría la prueba de hipótesis para proporciones para determinar la proporción de mejores soluciones encontradas mediante el uso de la metaheurística.
2. Explicar el valor de PIB en Colombia en función de variables como la tasa de interés, inversiones, número de empleados y tasas de cambio, utilizando un modelo de regresión lineal múltiple con la técnica de Mínimos Cuadrados Ordinarios (MCO).
3. Mediante información de la Encuesta Nacional de Niños, Niñas y Mujeres (ENIM) 2015 en México, se construye la variable binaria que tomará el valor de 1 si el niño (entre 0 y 5 años) está desnutrido y cero, en caso contrario. La desnutrición es medida a través del indicador talla para la edad. Se utilizarán variables explicativas como la edad de la madre, número de hermanos, región de

nacimiento y lactancia por parte de la madre, ingresos de la familia y, se utilizará un modelo de regresión Logit para tratar de explicar la probabilidad de si un niño es desnutrido o no.

4. Realizar un diseño de experimentos para evaluar entre un grupo de personas (por medio de una encuesta) la probabilidad estimada que tienen de ingresar a la Universidad considerando factores como la edad, genero, raza, ingresos mensuales, experiencia laboral, años de educación y si estan trabajando.

Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 16

1. Retroalimentación para propuestas del proyecto final de compañeros

1. **Propuesta 1 de Alberto Moa:** Es un interesante tema el cual te ayuda a enriquecer los resultados obtenidos y creo que estaría bien que trataras de implementar un diseño Taguchi y contrastar tus resultados.
2. **Propuesta 2 de Erick:** Interesante tema el cual ceo que podrías complementar utilizando grafos aleatorios y llevar a cabo una experimentación para analizar el efecto de agregar los nodos de transbordo.
3. **Propuesta 1 de Gabriela:** También podrías usar las pruebas de hipótesis de medias de diferencia para determinar que tanto mejoran las soluciones encontradas con los diferentes tipos de combinaciones.

Comparación de métodos de solución para la ubicación y recolección de productos en un almacén

Johanna Bolaños Zuñiga

Universidad Autónoma, San Nicolás de los Garza, Nuevo León México

Abstract

El proceso de preparación de pedidos es uno de los principales problemas en los almacenes, donde una de las actividades con mayor costo operativo es la recolección de pedidos. Esta actividad se logra mediante una política de enrutamiento que determina la secuencia que debe seguir el recolector de pedidos para tomar los artículos de las ubicaciones del almacén. Por lo tanto, las decisiones de asignación del espacio de almacenamiento influyen en la minimización del tiempo de recolección de pedidos y, en consecuencia, en la reducción de los costos de operación del almacén. De acuerdo a lo anterior, se cuenta con un modelo matemático y una metaheurística para determinar simultáneamente las decisiones de almacenamiento y rutas de recolección de los productos, considerando restricciones de precedencia con base en el peso del producto y las características del caso de estudio, como tener una única ubicación para cada producto en un almacén con un diseño general. En esta investigación mediante las pruebas de hipótesis sobre la media de la diferencia se resalta que entre los dos métodos propuestos estadísticamente la metaheurística presenta mejores resultados que la solución encontrada hasta el momento por el método exacto para la problemática planteada.

Keywords: Preparación de pedidos, Problema del recolector,
Almacenamiento, Prueba de hipótesis, Prueba *t*

1. Introducción

Según las necesidades de los clientes, el almacenamiento y el proceso de preparación de pedidos son un componente principal de cualquier cadena de

Email address: johana.bolanoszn@uanl.edu.mx (Johanna Bolaños Zuñiga)

suministro. Desde 1984 el Consejo de Educación e Investigación de Almacenaje (WERC por sus siglas en inglés de *Warehousing Education and Research Council*) [1] identificó que el proceso de preparación de pedidos (conocido como *order picking process*) es la principal área de oportunidad de la industria del almacenamiento. De acuerdo a estudios realizados por Tompkins et al. [2] y Henn and Schmid [3], es uno de los procesos más críticos a nivel operativo dentro de un almacén, debido a que representa entre el 55 % - 60 % de sus costos, razones por las cuales las empresas se ven obligadas a llevar a cabo esta actividad de la mejor manera posible.

Por otro lado, en la investigación de Goetschalckx and Ashayeri [4], el nivel de servicio de una empresa se compone de una variedad de factores tales como el promedio y la variación del tiempo de entrega de la demanda, la integridad y la precisión del producto. Por lo tanto, un vínculo crucial entre la preparación de pedidos y el nivel de servicio es que cuanto más rápido se realice la recolección de lo solicitado, más rápido estará disponible para enviarla al cliente. De lo contrario, es posible que se incurra en un atraso de la entrega provocando una mala prestación del servicio e o inconformidad por parte del cliente. No obstante, la empresa podría incurrir en trabajos adicionales para entregar a tiempo, elevando con ellos los costos de esta operación.

La preparación de pedidos implica una serie de actividades que van desde la selección o programación de los pedidos, la recolección de las cantidades de los diferentes artículos desde su ubicación de almacenamiento hasta el despacho de estos, en respuesta a las solicitudes de sus clientes. No obstante, para Davarzani and Norrman [5] y De Koster et al. [6] los objetivos a alcanzar en la recolección de pedidos son la minimización de las distancias o el tiempo de viaje que los recolectores realizan a través del almacén para cumplir con la demanda, actividades que se llevan a cabo por medio de políticas de enrutamiento, las cuales determinan la secuencia en la que el recolector de pedidos toma los artículos de las ubicaciones del almacén, y especifican tanto el orden en los que estos tienen que ser recogidos, así como el orden en la que se deben de visitar los pasillos.

Además, de acuerdo a las investigaciones de Brynzér and Johansson [7] y Manzini et al. [8] una correcta ubicación de almacenamiento de los productos facilita la precisión de proceso y la colocación más eficiente de las existencias, consiguiendo ciclos de pedidos más rápidos y con mejor servicio al cliente, lo que convierte al almacenamiento y el enrutamiento en uno de los principales problemas en la práctica. Para este trabajo se cuenta dos métodos de solución

(modelo matemático y una metaheurística) para resolver estos problemas de manera conjunta en una empresa donde no existe una ubicación apropiada de los productos, no se recolectan los productos con base a su peso y generan costos extras de personal para evitar incumplimiento de las entregas a los clientes. Se pretende comprobar que existe un ahorro entre la metaheurística y la solución ofrecida hasta el momento por el modelo matemático.

El resto de este trabajo está organizado de la siguiente manera: En la sección 2 se presenta una revisión de literatura que aborda los temas recolección de pedidos, almacenamiento, prueba de hipótesis de medias de diferencia. De igual manera, se mencionan los métodos utilizados para solucionar el problema presentado en este trabajo. En la sección 3 se describen las pruebas de hipótesis de medias de diferencia y de proporciones que serán utilizadas, mientras que la sección 4 se muestra el análisis y resultados propuestos en la sección anterior. Finalmente, la última sección se exponen las conclusiones, contribuciones y posible trabajo futuro de la investigación.

2. Antecedentes

De acuerdo a Daniels et al. [9], Lin et al. [10], Scholz et al. [11], Theys et al. [12], entre otros, la política de enrutamiento, puede interpretarse como un caso especial del *Travelling Salesman Problem* (TSP). Bajo este enfoque, en investigaciones como Petersen et al. [13], Ratliff and S. [14], Roodbergen and De Koster [15], Scholz et al. [11], Bolaños et al. [16] han aportado formulaciones para encontrar soluciones óptimas, en otras como Daniels et al. [9], Dekker et al. [17], De Koster and Poort [18], Theys et al. [12], Vaughan and Petersen [19], Žulj et al. [20], entre otros, han aportado soluciones heurísticas. Asimismo, Roodbergen and De Koster [15], Scholz et al. [11], Vaughan and Petersen [19] determinan que el rendimiento de estas estrategias (óptimas o heurísticas) depende de las características del problema, como lo son el tipo o tamaño del almacén, el número de pasillos de recolección, la cantidad de ubicaciones por pasillo y la ubicación del depósito.

Por otro lado, las investigaciones realizadas por De Koster et al. [6] y Bahrami et al. [21], mencionan que el método de solución más utilizado son los heurísticos puesto que son algoritmos se puede ajustar fácilmente a los cambios en el diseño y a las prioridades predeterminadas, sin embargo, no garantiza que sea la ruta o tiempo más corto posible.

Dicho lo anterior y de acuerdo con la extensa investigación realizada por Van Gils et al. [22], se establece que existen tres excelentes combinaciones

para mejorar la eficiencia de la preparación de pedidos, teniendo como mayor cantidad de casos de estudios el procesamiento por lotes y enrutamiento, seguido por la asignación de ubicación de almacenamiento y enrutamiento y por último la asignación de ubicación de almacenamiento y procesamiento por lotes. A pesar de que la mayor parte de la literatura se concentre en dar soluciones a la primera combinación, para Bartholdi and Hankman [23], el problema de enrutamiento del recolector es muy interdependiente del problema de asignación de almacenamiento. Cabe señalar que todos los estudios revisados por Van Gils et al. [22], los problemas fueron resueltos de manera independiente. Algunas de las investigaciones más relevantes que estudian los problemas de asignación de ubicación de almacenamiento y enrutamiento se pueden encontrar en Dekker et al. [17], Žulj et al. [20], Chabot et al. [24], Matusiak et al. [25], Daniels et al. [9], Bolaños et al. [16].

En gran parte de las investigaciones realizadas, la mayoría de los modelos o estrategias propuestas se basan en el cumplimiento de la demanda, dejando a un lado otros factores, como el peso de los productos, el cual es un criterio importante al momento de realizar la recolección, como se presenta en Dekker et al. [17], Žulj et al. [20], Chabot et al. [24], ya que conservar en buen estado el producto es de vital importancia para la satisfacción de los clientes, principalmente en almacenes donde se maneja productos frágiles. De acuerdo a esto, se hace importante el poder encontrar soluciones óptimas o mejores partiendo de la investigación realizada por Bolaños et al. [16], en la cual no se encuentran las soluciones óptimas para todas las instancias analizadas.

De acuerdo a lo anterior mencionado, se puede observar que se han propuesto tantos métodos óptimos como heurísticos para resolver el problema de enrutamiento del recolector y asignación de almacenamiento, por lo cual se hace interesante determinar cuál de las dos estrategias proporciona mejores resultados. Una manera de poder determinarlo es mediante las pruebas de hipótesis ya que es un procedimiento basado en evidencia muestral (estadístico) y en la teoría de probabilidad (distribución muestral del estadístico) para determinar si rechazar o no la hipótesis estadística acerca de una población.

En las pruebas mencionada anteriormente, se analizan dos hipótesis, la nula (H_0) y la alternativa (H_1). La primera es la afirmación que se está comprobando y, la segunda es una afirmación que se acepta si los datos muestrales proporcionan evidencia suficiente de que la hipótesis nula es falsa. Existen varias investigaciones como la de Saucedo [26] donde se emplean las pruebas de hipótesis para determinar estadísticamente la eficiencia de los tiempos de ejecución entre dos métodos de solución. Normalmente lo que se

hace es calcular un dato (media de las diferencias) que se compara con un estadístico de prueba y con base en ese estadístico se define que tan peor o mejor es una solución respecto a otra. Otras investigaciones donde utilizan las pruebas de hipótesis para hacer comparaciones de resultados se pueden encontrar en Puentes [27], Gao et al. [28].

3. Metodología

Como se expuso en la sección anterior, las pruebas de hipótesis se utilizan para hacer comparaciones. En este trabajo se utilizará la prueba de hipótesis de media de diferencias entre la solución encontrada hasta el momento por el modelo matemático y la metaheurística para comprobar que existe un ahorro y determinar que tanto mejora la solución.

Para la prueba de hipótesis de media de diferencias se utiliza la prueba t pareada ya que es una prueba robusta que se aplica con mayor frecuencia en problemas que implican muestras comparativas A. and S. [29], Johnston [30].

Como se cuenta con la solución encontrada hasta el momento por el método exacto y la metaheurística el problema de dos muestras se reduce en esencia a un problema de una muestra utilizando las diferencias calculadas entre el modelo matemático y la metaheurística d_1, d_2, \dots, d_n y el estadístico de prueba estará dado por la ecuación 1

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}, \quad (1)$$

donde \bar{d} es la media de las diferencias entre los dos métodos calculada o media muestral, μ_d es la media hipotética de las diferencias entre los dos métodos, s_d es la desviación estándar de las diferencias y n es el tamaño de la muestra. Los criterios o región crítica para rechazar la hipótesis nula H_0 planteada para esta prueba se muestra en el cuadro 1.

Una vez que se determine el ahorro entre la metaheurística y la solución ofrecida por el modelo matemático, se procede a aplicar la prueba de hipótesis para proporcionales para determinar la proporción de mejores soluciones encontradas mediante el uso de la metaheurística. En este tipo de prueba se considera el problema de probar la hipótesis de que la proporción de éxitos, fracasos, mejoras, etc., en un experimento binomial es igual a algún valor específico. Para este trabajo el estadístico de prueba estará dado por

Cuadro 1: Prueba de hipótesis para medias

Rechazar H_0 si:	
H_1	Criterio
$\mu_d > d$	$t > t_\alpha$
$\mu_d < d$	$t < -t_\alpha$
$\mu_d \neq d$	$ t > t_{\frac{\alpha}{2}}, n - 1$

la ecuación 2:

$$z = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1 - \theta_0)/n}}, \quad (2)$$

donde $\hat{\theta}$ es la proporción de mejoras observadas, θ_0 es la proporción de mejoras totales que se esperan tener y n es el tamaño de la muestra (Walpole et al. [31]). Los criterios o región crítica para rechazar la hipótesis nula H_0 . Los criterios o región crítica para rechazar la hipótesis nula H_0 planteada para esta prueba se muestra en el cuadro 2.

Cuadro 2: Prueba de hipótesis para proporciones

Rechazar H_0 si:	
H_1	Criterio
$\theta > \theta_0$	$z > z_\alpha$
$\theta < \theta_0$	$z < -z_\alpha$
$\theta \neq \theta_0$	$ z > z_{\frac{\alpha}{2}}$

3.1. Descripción de las instancias

Se cuenta con un total de 293 instancias analizadas, las cuales se dividen en 4 categorías: pequeñas, medianas tipo 1, medianas tipo 2 y grandes. El tipo de las instancias está relacionado con la cantidad de espacios disponibles para asignar producto y el tipo de productos solicitados en cada pedido. Las pruebas de hipótesis se aplican tanto por categoría como por el total de instancias y sólo se consideran aquellas en las que modelo matemático encontró una solución, quedando así una cantidad total de 185. Cabe mencionar que la metaheurística encontró soluciones a todas las 293 instancias.

4. Resultados y discusiones

En el cuadro 3, se muestra el número de instancias analizadas (n), la media muestral (\bar{d}) y desviación estándar (s_d) de las diferencias de las brechas (exacto - metaheurística / exacto) entre la metaheurística y el método exacto por categorías y por el total de instancias analizadas. La base de datos y el código en R utilizado, se encuentran disponibles en el repositorio de GitHub [32].

Cuadro 3: Parámetros estadísticos prueba de hipótesis para medias

Categorías	Media de diferencias \bar{d}	Desviación estándar s_d	Número de instancias n
Pequeñas	0.008	0.059	55
Medianas tipo 1	0.068	0.229	77
Medianas tipo 2	0.621	0.220	49
Grandes	0.677	0.170	4
Total instancias	0.210	0.326	185

4.1. Prueba de hipótesis de media de diferencias

La información del cuadro 3 se empleó para hacer la prueba de hipótesis de media de diferencias considerando un $\alpha = 0.05$, es decir, con un intervalo de confianza del 95 %. Se espera tener un ahorro (un tiempo de recolección mejor) entre la solución encontrada por la metaheurística y el modelo matemático. Este análisis se aplicó tanto por categoría como para el total de instancias. Se realizó el cálculo analítico con base en la ecuación 1 y como criterio de prueba para comparar las medias la información del cuadro 1. También se utilizó el programa R versión 4.0.2 [33] para aplicar la prueba t mediante la función `t.test`.

Instancias pequeñas

Sea $H_0: \mu_d = 0$ como hipótesis nula y como alternativa $H_1: \mu_d \neq 0$, es decir, se quiere comprobar que estadísticamente la metaheurística encuentra un mejor tiempo de recolección que la encontrada hasta el momento por el método exacto. De acuerdo al cuadro 1 el criterio de prueba que se utilizó para comparar las medias es el de $\mu_d \neq 0$, por lo tanto, los datos necesarios

para emplear el criterio son $n = 55$, $\bar{d} = 1.92$ y $s_d = 5.58$ (datos del cuadro 3). Reemplazando los datos anteriores en la ecuación 1 se tiene que:

$$t = \frac{0.0082 - 0}{0.0589/\sqrt{55}} = 1.038.$$

Ahora, el valor de $t_{\frac{0.05}{2}, 55-1} = 2.005$ (hallado con la función `qt`), por lo tanto, como $|1.038| < 2.005$ no rechazamos la H_0 lo que significa que, en promedio el tiempo de recolección encontrado por la metaheurística no es mejor que el reportado hasta el momento por el método exacto. Este resultado es posible ya que para este tipo de instancias la metaheurística logra alcanzar la misma solución del método exacto en 44 de las 55 instancias para esta categoría. Aplicando la función `t.test` en el programa R se puede observar que el valor $p > 0.05$.

Pruebas.txt

One Sample t-test

```
data: pequeñas
t = 1.0382, df = 54, p-value = 0.3038
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0.007677009 0.024167368
sample estimates:
mean of x
0.00824518
```

Instancias medianas tipo 1

Se repite el mismo procedimiento que se realizó para las instancias pequeñas, por lo tanto, los datos necesarios para emplear el criterio son $n = 55$, $\bar{d} = 1.92$ y $s_d = 5.58$ (datos del cuadro 3). Reemplazando los datos anteriores en la ecuación 1 se tiene que:

$$t = \frac{0.068 - 0}{0.229/\sqrt{77}} = 2.626.$$

Ahora, el valor de $t_{\frac{0.05}{2}}, 77 - 1 = 1.992$ (hallado con la función `qt`), por lo tanto, como $|2.626| > 1.992$ rechazamos la H_0 , por consiguiente, aceptamos la H_1 lo que significa que, en promedio el tiempo de recolección encontrado por la metaheurística es mejor que el reportado hasta el momento por el método exacto. Aplicando la función `t.test` en el programa R se puede observar que el valor $p < 0.05$

Pruebas.txt

One Sample t-test

```
data: medianas1
t = 2.626, df = 76, p-value = 0.01044
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.01653629 0.12037918
sample estimates:
mean of x
0.06845773
```

De acuerdo a lo anterior, podemos observar que hay una diferencia en el tiempo de recolección considerando el acomodo propuesto por la metaheurística y el del modelo matemático, ahora, se determina estadísticamente que tanto mejora la solución proporcionada por la metaheurística, para ello se considera una mejora del 95 % y 90 % o menor si fuera el caso.

Sea $H_0: \mu_d = 95 \% \bar{d}$ como hipótesis nula y como alternativa $H_1: \mu_d > 95 \% \bar{d}$, es decir que, se quiere comprobar que estadísticamente el tiempo de recolección encontrado por la metaheurística mejora en un 95 %.

De acuerdo al cuadro 1, el criterio de prueba que se utilizó para comparar las medias es el de $\mu_d > d$ (es decir, $t > t_\alpha$), por lo tanto, los datos necesarios para emplear el criterio son $n = 77$, $\bar{d} = 0.068$, $s_d = 0.229$ (datos del cuadro 3) y $\mu_d = 95 \% (0.068) = 0.065$. Reemplazando los datos anteriores en la ecuación 1 se tiene que:

$$t = \frac{0.068 - 0.065}{0.229/\sqrt{77}} = 0.131.$$

Ahora, el valor de $t_{0.05}, 77 - 1 = 1.665$ (hallado con la función `qt`), por lo tanto, como $0.131 \not> 1.665$ no se puede rechazar la H_0 , lo que significa que, en

promedio con un nivel de significancia de $\alpha = 0.05$ el tiempo de recolección encontrado por metaheurística no mejora en un 90 % al reportado por el método exacto, por lo tanto, se prueba con un porcentaje menor de mejora, para este caso se considera un 35 %,

Calculando nuevamente $\mu_d = 35\%(0.068) = 0.024$ se obtiene un valor del estadístico $t = 2.037$, por lo tanto, como $1.707 > 1.674$ se rechaza la H_0 , lo que significa que, en promedio el tiempo de recolección encontrado por metaheurística mejora en un 35 % al reportado por el método exacto.

Pruebas.txt

One Sample t-test

```
data: mejoramedianas1
t = 1.7069, df = 76, p-value = 0.04596
alternative hypothesis: true mean is greater than 0.02396021
95 percent confidence interval:
0.02504846      Inf
sample estimates:
mean of x
0.06845773
```

El anterior procedimiento se repite para las categorías medianas tipo 2 y grandes y para las instancias totales, los resultados se consolidan en el cuadro 6.

Cuadro 4: Resultados prueba de hipótesis para comprobar si la metaheurística mejora las soluciones

Categoría	Prueba de hipótesis	Criterio de rechazo $ t > t_{\frac{\alpha}{2}, n-1}$	Rechazo de la H_0
Mediana tipo 2	$H_0 : \mu = 0$	$ 19.774 > 2.011$	Si
Grandes	$H_1 : \mu \neq 0$	$ 7.944 > 3.182$	Si
Total instancias		$ 8.776 > 1.973$	Si

De acuerdo al cuadro 6, se puede determinar estadísticamente que tanto para las instancias medianas tipo 2, grandes y el total, en promedio el tiempo de recolección encontrado por la metaheurística es mejor que el reportado hasta el momento por el método exacto. Aplicando la función `t.test` en el programa R se puede observar que el valor $p < 0.05$.

Pruebas.txt

One Sample t-test

```
data: medianas2
t = 19.774, df = 48, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.5579331 0.6842349
sample estimates:
mean of x
 0.621084
```

Pruebas.txt

One Sample t-test

```
data: grandes
t = 7.9439, df = 3, p-value = 0.00416
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4058593 0.9483968
sample estimates:
mean of x
 0.677128
```

Pruebas.txt

One Sample t-test

```
data: totalinstancias
t = 8.7762, df = 184, p-value = 1.142e-15
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1628593 0.2573175
sample estimates:
mean of x
 0.2100884
```

A continuación, en el cuadro 5 se presentan los resultados obtenidos para determinar en qué porcentaje se mejoran las soluciones encontradas por la metaheurística.

De acuerdo al cuadro 5, se puede determinar estadísticamente que en promedio el tiempo de recolección encontrado por la metaheurística mejora

Cuadro 5: Resultados prueba de hipótesis sobre el porcentaje de mejora de la solución encontrada por la metaheurística

Categoría	Prueba de hipótesis	Criterio de rechazo $t > t_\alpha$	Rechazo de la H_0
Mediana tipo 2	$H_0 : \mu_d = 95\% \bar{d}$ $H_1 : \mu_d > 95\% \bar{d}$	$0.988 \not> 1.677$	No
	$H_0 : \mu_d = 90\% \bar{d}$ $H_1 : \mu_d > 90\% \bar{d}$	$1.977 > 1.677$	Si
Grandes	$H_0 : \mu_d = 95\% \bar{d}$ $H_1 : \mu_d > 95\% \bar{d}$	$0.397 \not> 2.353$	No
	$H_0 : \mu_d = 90\% \bar{d}$ $H_1 : \mu_d > 90\% \bar{d}$	$0.794 \not> 2.353$	No
	$H_0 : \mu_d = 70\% \bar{d}$ $H_1 : \mu_d > 70\% \bar{d}$	$2.383 > 2.353$	Si
Total instancias	$H_0 : \mu_d = 95\% \bar{d}$ $H_1 : \mu_d > 95\% \bar{d}$	$0.438 \not> 1.653$	No
	$H_0 : \mu_d = 90\% \bar{d}$ $H_1 : \mu_d > 90\% \bar{d}$	$0.878 \not> 1.653$	No
	$H_0 : \mu_d = 80\% \bar{d}$ $H_1 : \mu_d > 80\% \bar{d}$	$1.755 > 1.653$	Si

las soluciones reportadas hasta el momento por el método exacto en un 90 % las instancias medianas tipo 2, en un 70 % las grandes y en un 80 % el total de estas.

Con base en los resultados mostrados anteriormente, se demuestra estadísticamente que la metaheurística presenta mejores resultados en el tiempo de recolección para las instancias donde el modelo matemático presenta una mayor complejidad, como lo son las categorías medianas y grandes, lo cual es un comportamiento que se espera obtener cuando se plantea un algoritmo metaheurístico para resolver el tipo de problema presentado en este trabajo. En la figura 1 se puede observar el comportamiento del porcentaje de mejora mencionado.

Ahora, por medio de la prueba de hipótesis para proporciones se determinará estadísticamente la cantidad de instancias en las la metaheurística encuentra una mejor solución. Esta prueba se realizará para el total de las instancias en las que el modelo matemático reportó una solución y por la categoría mediana (tipo 1 y 2). Las pequeñas ni las grandes se consideran ya que

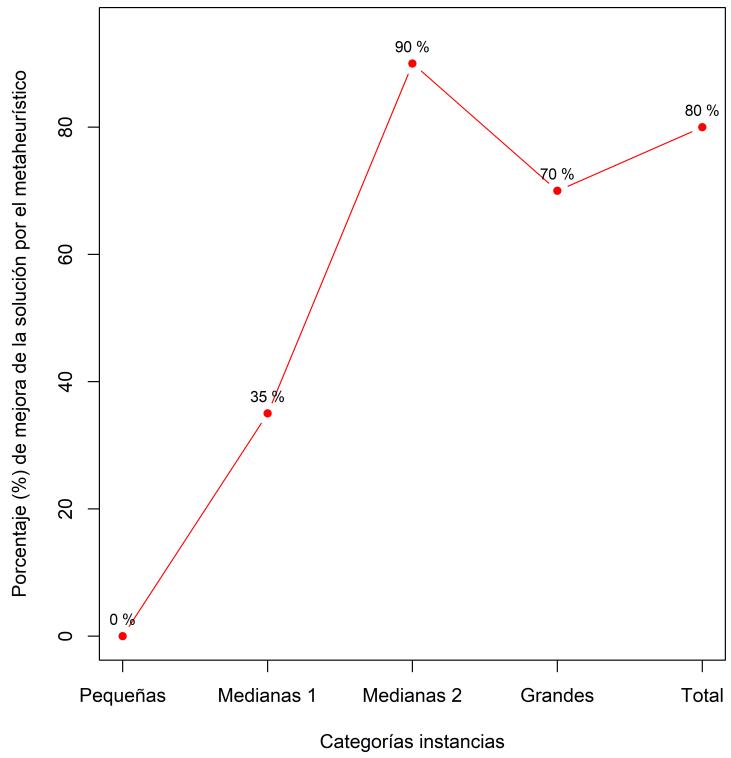


Figura 1: Comportamiento del porcentaje de mejora de la solución encontrada por la metaheurística por categorías de instancias

par a las primeras la prueba t mostró que no hay mejora y para las segundas la solución de todas las instancias fue mejoradas por la metaheurística.

4.2. Prueba de hipótesis para proporciones

La información del cuadro 2 se empleó para hacer la prueba de hipótesis para proporciones considerando un $\alpha = 0.05$, es decir, con un intervalo de confianza del 95 %. Se espera que del total de los casos (por categoría y global) un mínimo del 60 % sean detectados como mejoras. Cabe mencionar que para los casos donde la metaheurística encontró la misma solución que el modelo, se tomará como mejora ya que el tiempo de ejecución de la metaheurística es considerablemente menor que el requerido por el modelo para encontrar una solución.

Cuadro 6: Parámetros prueba de hipótesis para proporcioness

Categorías	Número de instancias n	Instancias observadas con mejora x	Proporción de instancias con mejora $\hat{\theta} = x/n$
Medianas tipo 1	77	67	0.870
Medianas tipo 2	49	48	0.980
Total instancias	185	173	0.935

Se realizó el cálculo analítico con base en la ecuación 2 y como criterio de prueba para las proporciones la información del cuadro 2. También se utilizó el programa R para aplicar esta prueba mediante la función `prop.test`.

Instancias medianas tipo 1

Sea $H_0: \theta = \theta_0$ como hipótesis nula y como alternativa $H_1 : \theta > \theta_0$, es decir, se demostrará estadísticamente que la metaheurística cumple con los cantidad de casos esperados, por lo tanto, los datos necesarios para emplear el criterio son $n = 77$, $\theta_0 = 0.6$ y $\hat{\theta} = 0.870$ (datos del cuadro 3. Reemplazando los datos anteriores en la ecuación 2 se tiene que:

$$t = \frac{0.870 - 0.60}{\sqrt{(0.60(1 - 0.6))/77}} = 4.839.$$

Ahora, el valor de $z_\alpha = 1.645$ (valor más exacto de acuerdo a la tabla de distribución z consultada en Walpole et al. [31]), por lo tanto, como $4.839 > 1.645$ rechazamos la H_0 lo que significa que la proporción de instancias mejoradas por la metaheurística en las instancias medianas tipo 1 es de por lo menos un 60 %. Aplicando la función `prop.test` en el programa R se puede observar que el valor $p < 0.05$.

————— Pruebas.txt —————

1-sample proportions test without continuity correction

```
data: 67 out of 77, null probability 0.6
X-squared = 23.411, df = 1, p-value = 6.541e-07
alternative hypothesis: true p is greater than 0.6
95 percent confidence interval:
```

```

0.7943705 1.0000000
sample estimates:
p
0.8701299

```

El anterior procedimiento se repite para las categorías medianas tipo 2 y para el total de las instancias, los resultados se consolidan en el cuadro 7.

Cuadro 7: Resultados prueba de hipótesis para las proporciones

Categorías	Prueba de hipótesis	Criterio de rechazo $z > z_{alpha}$	Rechazo de la H_0
Medianas tipo 2	$H_0 : \theta = \theta_0$	5.424 > 1.645	Si
Total instancias	$H_1 : \theta > \theta_0$	9.304 > 1.645	Si

De acuerdo al cuadro 7, se puede determinar estadísticamente que tanto para la categoría medianas tipo 2 como para el total de instancias, la proporción de instancias mejoradas por la metaheurística es de por lo menos un 60 %.

5. Conclusiones

Las decisiones de la ubicación de almacenamiento de los productos, son de vital importancia para determinar su facilidad de acceso, por lo tanto, contar con una herramienta que permita obtener un correcto acomodo, con base en criterios como el peso de los productos, cumplimiento de demanda, entre otros, es uno de los factores principales que permite realizar la recolección de pedidos en tiempo y forma y con ello mantener un nivel de respuesta rápida a la solicitud de los clientes.

Con las pruebas de hipótesis de medias de diferencia y de proporciones se comprueba estadísticamente que la metaheurística propuesta para solucionar de manera simultánea la ubicación de almacenamiento y las rutas de recolección con base a la prioridad de peso de las cajas de los productos y cumplimiento de demanda presenta mejores resultados que las soluciones reportadas hasta el momento por el modelo matemático, principalmente en las instancias que presentaron una mayor complejidad computacional.

Como trabajo futuro ya que se comprobó estadísticamente que las metaheurística mejora las soluciones encontradas hasta el momento por el modelo,

sería interesante que estas soluciones fueran soluciones iniciales del método exacto en las instancias de altos complejos para mejorar el tiempo computacional y encontrar soluciones óptimas al problema y mediante el uso de las pruebas de hipótesis comprobar si la combinación de estos métodos aportan mejores soluciones que las encontradas hasta el momento para este tipo de problemas.

6. Agradecimientos

Al Consejo de Ciencia y Tecnología (CONACYT) y a la Universidad Autónoma de Nuevo León por apoyarme con la beca que me permite realizar mis estudios, a la Dra. Elisa Schaeffer por sus enseñanzas y conocimientos transmitidos durante la clase de modelos probabilísticos aplicados (semestre agosto 2020 - enero 2021) y a las Dra. María Angélica Salazar Aguilar y Dra. Jania Astrid Saucedo Martínez por su apoyo y asesoría durante este proyecto.

Referencias

- [1] Goetschalckx, M.; Ratliff, H. Order Picking In An Aisle. *1988*, *20*, 53–62.
- [2] Tompkins, J.; J., W.; Y., B.; A., T. J. M. *Facilities planning 4th ed*; John Wiley and Sons, 2010.
- [3] Henn, S.; Schmid, V. Metaheuristics for order batching and sequencing in manual order picking systems. *Computers and Industrial Engineering* **2013**, *66*.
- [4] Goetschalckx, M.; Ashayeri, J. Classification and design of order picking systems. *Logistics World* **1989**, *2*, 99—106.
- [5] Davarzani, H.; Norrman, A. Toward a relevant agenda for warehousing research: literature review and practitioners' input. *Logistics Research* **2016**, *8*.
- [6] De Koster, R.; Le-Duc, T.; Roodbergen, K. J. Design and control of warehouse order picking: A literature review. *European Journal of Operational Research* **2007**, *182*.

- [7] Brynzér, H.; Johansson, M. Storage location assignment: Using the product structure to reduce order picking times. *International Journal of Production Economics* **1996**, *46-47*, 595–603.
- [8] Manzini, R.; Bindi, F.; Ferrari, E.; Pareschi, A. *Correlated Storage Assignment and Iso-Time Mapping Adopting Tri-Later Stackers. A Case Study from Tile Industry*; Springer, London, 2012; pp 373–396.
- [9] Daniels, R. L.; Rummel, J. L.; Schantz, R. A model for warehouse order picking. *European Journal of Operational Research* **1998**, *105*, 1–17.
- [10] Lin, C.-C.; Kang, J.-R.; Hou, C.-C.; Cheng, C.-Y. Joint order batching and picker Manhattan routing problem. *Computers and Industrial Engineering* **2016**, *95*, 164–174.
- [11] Scholz, A.; Henn, S.; Stuhlmann, M.; Wäscher, G. A new mathematical programming formulation for the Single-Picker Routing Problem. *European Journal of Operational Research* **2016**, *253*.
- [12] Theys, C.; Brásy, O.; Dullaert, W.; Raa, B. Using a TSP heuristic for routing order pickers in warehouses. *European Journal of Operational Research* **2010**, *200*.
- [13] Petersen, C. G.; Schmenner, I.-I.; Roger, W. An Evaluation of Routing and Volume-based Storage Policies in an Order Picking Operation. *Computers & Operations Research* **1999**, *30*, 481–501.
- [14] Ratliff, H. D.; S., R. A. Order-Picking in a Rectangular Warehouse: A Solvable Case of the Traveling Salesman Problem. *Operations Research* **1983**, *31*, 507–521.
- [15] Roodbergen, K. J.; De Koster, R. B. M. Routing methods for warehouses with multiple cross aisles. *International Journal of Production Research* **2001**, *39*, 1865–1883.
- [16] Bolaños, J.; Saucedo, J. A.; Salais, T. E.; Marmolejo, J. A. Optimization of the storage location assignment and the picker-routing problem by using mathematical programming. *Applied Science* **2020**, *10*, 534.
- [17] Dekker, R.; De Koster, R. B. M.; Roodbergen, K. J.; Van, K. H. Improving Order Picking Response Time at Ankor's Warehouse. *Interfaces* **2004**, *34*, 303–313.

- [18] De Koster, R.; Poort, E. V. D. Routing orderpickers in a warehouse: a comparison between optimal and heuristic solutions. *IIE Transactions* **1998**, *30*, 469–480.
- [19] Vaughan, T.; Petersen, C. The effect of warehouse cross aisles on order picking efficiency. **1999**, *37*, 881–897.
- [20] Žulj, I.; Glock, C. H.; Grosse, E. H.; Schneider, M. Picker routing and storage-assignment strategies for precedence-constrained order picking. *Computers Industrial Engineering* **2018**, *123*, 338–347.
- [21] Bahrami, B.; Aghezzaf, E.; Limere, V. Using Simulation to Analyze Picker Blocking in Manual Order Picking Systems. *Procedia Manufacturing* **2017**, *11*, 1798–1808, 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy.
- [22] Van Gils, T.; Ramaekers, K.; Caris, A.; De Koster, R. B. M. Designing efficient order picking systems by combining planning problems: State-of-the-art classification and review. *European Journal of Operational Research* **2018**, *267*, 1–15.
- [23] Bartholdi, J.; Hankman, S.
- [24] Chabot, T.; Lahyani, R.; Coelho, L. C.; Renaud, J. Order picking problems under weight, fragility and category constraints. *International Journal of Production Research* **2017**, *55*, 6361–6379.
- [25] Matusiak, M.; De Koster, R.; Kroon, L.; Saarinen, J. A fast simulated annealing method for batching precedence-constrained customer orders in a warehouse. *European Journal of Operational Research* **2014**, *236*, 968–977.
- [26] Saucedo, J. A. Verificación y empleo computacional de un modelo matemático utilizado para un layout en un centro de distribución. Ph.D. thesis, Universidad Autónoma de Nuevo León, 2005.
- [27] Puentes, D. E. Diseño de un modelo de distribución óptimo para un área de almacenamiento de operación manual basado en la estrategia “Forward reserve”. Ph.D. thesis, Universidad Autónoma de Nuevo León, 2016.

- [28] Gao, J.; Wu, Y.; Shen, T. Experimental comparisons of hypothesis test and moving average based combustion phase controllers. *ISA Transactions* **2016**, *65*, 504–515.
- [29] A., A. K.; S., P. Making Sense of Methods and Measurement: t Test Part II. *Clinical Simulation in Nursing* **2014**, *10*, e223.
- [30] Johnston, L. W. Student's t-Test. *Journal of Quality Technology* **1970**, *2*, 243–245.
- [31] Walpole, R. E.; Myers, R. H.; Myers, S. L.; Ye, K. *Probability statistics for Engineers Scientists*, 9th ed.; Pearson Education, Inc., 2012.
- [32] Bolaños, J., Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/proyecto, 2020.
- [33] The R Foundation, The R Project for Statistical Computing. <https://www.r-project.org/>, 2020.